REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
29-08-2012

2. REPORT TYPE
FINAL

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
Human-Robot Teams Informed by Human Performance Moderator Functions

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-09-1-0108

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)
Julie A. Adams, Ph.D., Vanderbilt University
Scott A. DeLoach, Ph.D., Kansas State University

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Vanderbilt University
2301 Vanderbilt Place
Nashville, TN 37235-1824

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR
875 N RANDOLPH ST
ARLINGTON, VA 22203
Dr. Robert Bonneau/RSL

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-1176

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This research resulted in several key results. These results are highlighted here and discussed in more detail in the following sections.
Teaming scenarios were modeled using IMPRINT Pro in order to predict the impact of various HPMFs on the teaming relationship. The developed human-human and human-robot models represented teaming relationships in which a) one team member was guided through a series of task steps and b) collaborative teams participate in tasks incorporating joint decision making for environments involving higher levels of uncertainty. These models provide predictions of human performance when working with a human or robot partner, and how performance differs based on the particular teaming partner. Model evaluations were conducted with single and multiple HPMFs and informed the validation evaluation designs. Additionally, the modeling results can inform human-robot teaming allocations.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
U	U	U	U		Julie A. Adams, Ph.D.

19b. TELEPHONE NUMBER *(Include area code)*
Phone: (615) 322 – 8481

Reset

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

FINAL REPORT

Human-Robot Teams Informed by Human Performance Moderator Functions

Grant Number: FA9550-09-1-0108

August 29, 2012

Julie A. Adams, Ph.D.	**Scott A. DeLoach, Ph.D.**
Vanderbilt University	Kansas State University
2301 Vanderbilt Place	234 Nichols Hall
Nashville, TN 37235-1824	Manhattan, KS 66506-2302
Phone: (615) 322 – 8481	Phone: (785) 532 – 6350
Fax: (615) 343 -5459	Fax: (785) 532 – 7353
E-mail: julie.a.adams@vanderbilt.edu	E-mail: sdeloach@cis.ksu.edu

1 Introduction

Rising expectations for adaptive computing systems include the ability to automatically correct themselves in a fluid and dynamic environment (e.g., autonomic systems [42]). Multiagent concepts [14] are well-suited for developing adaptive systems. Russell and Norvig [62] define agents as able to perceive and act autonomously such that their actions are based on their own experiences rather than predefined knowledge. Multiagent Systems (MASs) exploit this behavior to self-correct; if one agent should fail, another agent can take over.

One approach in multiagent research is to leverage organizational concepts such as agents, roles, and goals found in organizational models to produce Organization-based Multiagent System (OMAS). Some examples of such organizational models are Organization Model for Adaptive Computational Systems (OMACS) [19], Organizational Model for Normative Institutions (OMNI) [22], Organizations per Agents (OperA) [21], and HarmonIA [84]. By leveraging these organizational models, a general approach to adaptivity can be achieved through task allocations. Task allocations can be handled in a general manner because these models capture the necessary information to reallocate a task should an agent fail. Various research teams have applied organizational concepts in robotics, particularly in multirobot systems [12, 26, 28, 55, 69, 72, 76].

Another way of increasing a system's ability to adapt is by including humans as part of the system. Traditionally, humans have been considered as users of a computing system; humans are not typically considered as a factor during a system's decision making process. As computing systems continue to grow, the environments in which these systems operate inevitably involve humans. By including humans as a factor in these systems' decision making process, such systems are able to increase their adaptivity; tasks that cannot be completed by the system due to failures can be allocated to humans for completion and vice versa. There are two aspects involved when attempting to include humans as part of a system. First, designers must consider an interface to allow humans to interact with the system and vice versa. The actual requirements for such interfaces are beyond the scope of this report. Second, an appropriate internal structure for a system to support humans so that the system can reason about humans and their abilities to complete tasks must be developed. This aspect significantly increases the complexity of such systems. One way to mitigate the increase in complexity is to represent humans in a general manner such that systems can reason about humans in an abstract way. Fortunately, organization-based models are well-suited to facilitate integration of humans because these models already provide a basic framework for representing humans; humans can be considered as agents. This leads to one of the main questions addressed in the research: *what type of information about the humans should be captured that can lead to better allocation of tasks*.

Some investigation of the types of information about humans relevant to task allocation already exists. One type of information that is relevant for task allocation is human performance factors[1]. Wickens et al. [87] examined and explained a large variety of human performance factors that are relevant to designing systems

[1] The term "human factors" has multiple meanings. In order to distinguish between them, the term "human performance factors" refers to a specific definition where human factors are factors that affect the performance of an individual.

for human interaction. For instance, they explain the various human performance factors that affect the ability of a human to drive at night, which includes the eyesight of the driver, the fatigue level of the driver, the reaction time of the driver, the color of objects, the luminosity of objects, the current weather conditions, the ambient lighting, and the speed of the car. For example, suppose there is one last task to deliver a package, it is snowing heavily, and there are two drivers available. Driver A has been driving for the past eight hours and is fatigued; however, driver B has only been driving for four hours and is less fatigued than driver A. Thus, it is better to pick driver B who is less fatigued for the task. These human performance factors are not exclusive to any particular task and they can be classified into three categories: human-specific, task-specific, and environment-specific. However, there are a number of challenges to overcome in order to use human performance factors.

- Define a means of capturing human performance factors so that they can be used by task allocation algorithms.
- Define an appropriate mechanism so that a large number of human performance factors can be used at runtime for computing task allocations.

Performance moderator functions [64] have been used to capture human performance factors. Specifically, Human Performance Moderator Functions (HPMFs) have been used to capture human-specific human performance factors. Performance moderator functions are a well-known and accepted approach to capturing human performance factors. In general, there are an enormous number of human performance factors [87]. Thus, when designing computing systems that include humans as part of the system (i.e., humans are considered as peers), there can be a significant increase in the amount of information to be handled and thus, the complexity of these systems can become overwhelming [64]. This leads to the need for an appropriate mechanism such that the complexity of systems that includes humans is not overwhelming. The complexities of including HPMFs can be managed by leveraging Model Driven Engineering [34]. By following the model driven engineering approach, a runtime model (commonly referred to as *models@run.time* [9]) can be developed. This runtime model allows development of an adaptive mechanism that can autonomously perform task allocations. Furthermore, in the field of autonomous task allocation for multirobot systems, Parker [56] identified three paradigms for tackling the problem of task allocation. Two of the paradigms (the role-based organizational paradigm and the knowledge-based paradigm) tackle the problem of task allocations for heterogeneous robots in different ways. We followed the approach of OMACS, which combines both paradigms. However, the problem of task allocation is NP-hard [29], and thus, it is not realistic to expect optimal task allocations during runtime as general optimal task allocation algorithms take too much time. In practical terms, greedy-based task allocation algorithms are often "good enough". Thus, the approach does not assume optimal task allocations from the algorithms.

A realistic application of this research is in the allocation of humans and robots to teams for complex missions that require peer-based interaction [32, 65]. Thus, the presented research had two primary focuses. The first focus was on analyzing and validating HPMFs for a human-robot team in two team configurations,

one that required the human to carry out instructions provided by the robot (or, for purposes of validation, human) partner and another that required the team to collaborate and make joint decisions. The second research focus was on integrating the consideration of HPMFs into the multiagent task allocation framework and analyzing the ability to allocate tasks across the team members. The first research focus reviewed a broad spectrum of HPMFs, modeled a subset of the reviewed HPMFs and conducted validation evaluations for workload and reaction time. The second focus developed a computational framework for determining the effect of HPMFs on human abilities to perform team roles and adjusting task allocations based on the expected human performance.

1.1 Key Results

This research resulted in several key results. These results are highlighted here and discussed in more detail in the following sections.

Teaming scenarios were modeled using IMPRINT Pro in order to predict the impact of various HPMFs on the teaming relationship. The developed human-human and human-robot models represented teaming relationships in which a) one team member was guided through a series of task steps and b) collaborative teams participate in tasks incorporating joint decision making for environments involving higher levels of uncertainty. These models provide predictions of human performance when working with a human or robot partner, and how performance differs based on the particular teaming partner. Model evaluations were conducted with single and multiple HPMFs and informed the validation evaluation designs. Additionally, the modeling results can inform human-robot teaming allocations. Evaluations incorporating the human-human and human-robot teaming relationships were conducted in order to validate the models of workload and reaction time.

The human-robot teams resulted in lower workload, independent of teaming relationship and the IMPRINT Pro models provided a good prediction of workload independent of the teaming relationships. It was also determined that neither the teaming relationship nor the teaming partner (human or robot) negatively impacted task performance. Due to technology limitations, the human-robot teams required more time to complete the tasks than the human-human teams, independent of teaming relationship. The extended task completion times for the human-robot teams were not the sole influence on lowering workload, but the additional factors that contributed to reducing workload differed by teaming relationship.

Methods of measuring reaction time during user evaluations that occur when interacting with the real-world, rather than sitting in front of a computer, were developed. The methods included using a wearable point-of-view camera to determine visual fixation, a laser pointer to point at the item of interest, verbal utterances about an object in the environment, tactile interaction with an object or taking a picture of an object. Visual fixation and verbal utterances were the most frequently used identifiers. This research represents a preliminary step in identifying new and reliable measures of reaction time outside of constrained laboratory interaction settings.

The IMPRINT Pro model for collaborative teams was a good predictor of reaction time for the human-human teams, but severely underestimated the human's reaction time when partnered with a robot. A number of theories are presented as to the cause of the slower reaction times during the human-robot teaming that will be the focus of future research.

New stressors were developed and integrated into IMPRINT Pro in order to understand the impact of reaction time and vigilance. A new time-based stressor for vigilance was developed based on theories related to the decrease in human signal detection over time. A new IMPRINT Pro micromodel for reaction time was developed based on the results from the collaborative team validation evaluation that accounts for visual system recognition. The existing reaction time micromodel underestimated the human's reaction time when teamed with a robot.

The Chazm model (CzM) was created to incorporate HPMFs into a framework amenable to computation of the effect of task enactment on human performance over time. Specifically, the model captured the notion of tasks, performance functions, and human performance attributes. Several algorithms were created that used the CzM model to compute the expected performance degradations based on current and future task allocations. In addition, the algorithms modified the allocation of tasks based on the human performance attribute values.

The usefulness of the CzM model was also evaluated by applying the related algorithms in several different simulated experiments. These simulations modeled the task allocation process in purely human applications as well as combined human-robot teams. The results of the evaluation showed that CzM can capture HPMFs and use their results it runtime to modify task allocations in order to reduce overall human workload, while still accomplishing all team tasks.

1.2 Publications

Caroline E. Harriott, Tao Zhang and Julie A. Adams. Assessing Physical Workload for Human-Robot Peer-based Teams. International Journal of Human-Computer Systems (In Revision).

Caroline E. Harriott, Glenn L. Buford, Tao Zhang, and Julie A. Adams. Workload Prediction for Collaborative Human-Robot Teams. Journal of Human-Robot Interaction (In Preparation).

Caroline E. Harriott, Glenn L. Buford, Tao Zhang, and Julie A. Adams. Prediction and Assessment of Reaction Time for Human-Robot Teams. ACM/IEEE International Conference on Human-Robot Interaction (In Preparation).

Caroline E. Harriott, Glenn L. Buford, Tao Zhang, and Julie A. Adams. Assessing Workload in Human-Robot Peer-Based Teams in Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction, 2012.

Caroline E. Harriott, Tao Zhang and Julie A. Adams. Predicting and validating workload in human-robot teams in Proceedings of the 20th Conference on Behavior Representation in Modeling and Simulation, pages 162-169, 2011.

Caroline E. Harriott, Tao Zhang and Julie A. Adams. Evaluating the applicability of current models of workload to peer-based human-robot teams. In Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction, pages 45-52, 2011.

Caroline Harriott, Tao Zhang and Julie A. Adams. Applying workload human performance moderator functions to peer-based human-robot interaction. Technical report HMT-10-04, *Human-Machine Laboratory Technical Report*, 2010.

Christopher Zhong and Scott A. DeLoach. Using Performance Moderator Functions to Model Humans in Automatic Task Allocation for Human-Robot Teams. Journal of Autonomous Agents and Multiagent Systems. (*submitted*)

Caroline E. Harriott, Rui Zhuang, Julie A. Adams and Scott A. DeLoach. Towards Using Human Performance Moderator Functions in Human-Robot Teams. First International Workshop on Human-Agent Interaction Design and Models (HAIDM 2012). Valencia, Spain, June 4, 2012.

Christopher Zhong and Scott A. DeLoach. Runtime Models for Automatic Reorganization of Multi-Robot Systems. 6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS 2011). Waikiki, Honolulu, Hawaii, May 23-24, 2011.

Christopher Zhong and Scott A. DeLoach. Integrating Performance Factors into an Organization Model for Multiagent Systems. Multiagent & Cooperative Robotics Laboratory Technical Report No. MACR-TR-2010-05. Kansas State University.

Christopher Zhong and Scott A. DeLoach. Integrating Performance Factors into an Organization Model for Better Task Allocation in Multiagent Systems. Multiagent & Cooperative Robotics Laboratory Technical Report No. MACR-TR-2010-02. Kansas State.

1.3 Overview

The rest of this report is organized as follows. Section 2 highlights the foundational work on which this research is based. Our modeling and validation evaluations of workload for Guided Interaction Teams are presented in Section 3, while the research related to workload and reaction time for Collaborative Teams is presented in Section 4. Section 5 presents results from modeling additional human performance moderator functions. Section 6 describes our runtime model that integrates HPMF into a runtime framework for use in task allocation. Finally, Section 7 provides a conclusion.

2 Background

2.1 Human Performance Moderator Functions (HPMFs)

The study of human factors has led to the development of performance moderator functions. The goal of human factors (as defined by Wickens et al. [87]) is to facilitate better human interaction with systems so that (1) performance is enhanced, (2) safety is increased, and (3) user satisfaction is increased.

Human factors play an important role in the development of simulations that model human behavioral and cognitive processes. For example, the Department of Defense (DoD) Modeling & Simulation Coordination Office (M&S CO) uses human factors to create realistic and complex virtual worlds for training soldiers. However, the literature in human factors is vast and extracting these factors into a form that is suitable for implementation often requires expertise in the associated field. HPMFs are a way to bridge the gap between literature and implementation.

HPMFs [66] indicate the impact of internal and external human factors on human performance. Examples of internal human factors are fatigue level, reaction time, and mental acuity. Examples of external human factors are noise level, lighting, and task time. In addition, HPMFs are able to capture impact of personality on performance such as emotion, cultural background, and biases. Furthermore, HPMFs quantify performance differences between two humans such as intelligence, skill, and motivation. In other words, HPMFs capture the relationship between performance moderators and the level of performance in the form of dose-response (or exposure-response). As the dose (or exposure) increases or decreases, HPMFs indicate the change in the level of performance.

Human performance modeling simulates human behavior under various conditions and tasks. The models require inputs related to human performance and result in the likely actions. HPMFs can be incorporated into human performance models in order to improve the model fidelity [1]. A large number of domains have incorporated human performance models in order to understand how system design, task assignments and environmental changes can impact human behavior and performance. For example, in the aviation field, the NASA Human Performance Modeling Project [25] involved using multiple modeling techniques on a common set of problems to investigate different aspects of human performance in aviation tasks. The effect of a secondary task on human performance while driving a car was analyzed [63] using ACT-R [4].

PMFServ represents human decision-making processes based on a subjective utility ranking [68]. PMFServ incorporates the effects of HPMFs to affect each agent's decision-making process. PMFServ has been used to model human behavior in many scenarios including hostile civilians in an urban military scene [79].

IMPRINT Pro is a task network modeling tool intended to assess human and system performance in military missions [3, 5]. IMPRNT Pro has been used to model personnel on a United States Navy destroyer bridge, the U.S. Army's Crusader System [1], and pilot performance for simulated unmanned air vehicles missions [86]. IMPRINT Pro simulates human behavior for a variety of conditions through the representation of task and event networks. IMPRINT Pro includes a number of pre-defined HPMFs (e.g., workload) and permits

the incorporation of undefined HPMFs via the User Stressors module. IMPRINT Pro has been employed in the reported research to model both human-human and human-robot teams.

The reported research began with a literature review to identify human performance moderator functions that were deemed appropriate for the research project and applicable to human-robot peer-based teams. The literature review identified a total 58 human performance moderator functions that were identified by category: Visual (e.g. visibility, visual selective attention, etc.), Auditory (e.g., noise level, etc.), Cognitive (e.g., stress, vigilance, workload, etc.), Psychomotor (e.g., temperature, vibration, etc.), Organizational (e.g., rank within organization, policies and procedures, etc.), Ergonomic (e.g., health and illness, physical size, etc.), Social (Cultural values and racial attitude). These 58 HPMFs were narrowed to 20 HPMFs that were realistic for consideration with human-robot teams and conducive for analysis.

The 20 HPMFs were organized based on four characteristics: Acquired, Environmental, Designed, and Dynamic. Acquired HPMFs change slowly over time and are generally under the control of the individual human. Environmental HPMFs are dynamic in nature, not controlled by humans, generally are a constant factor, but may change slowly over time. Designed HPMFs are not under the control of the human and cannot be consciously improved or degraded over time, rather they represent inherent abilities. Dynamic HPMFs can change quickly and are influenced by the situation. Table 1 provides the HPMFs organized by these categories. It should be noted that some HPMFs can be in two categories, but one category is deemed the primary category. The category that contains the shaded HPMF represents the secondary category.

Table 1. The twenty considered HPMFs organized by category. HPMFs with gray background appear in two categories and those with gray background represent the secondary categorization.

Acquired	Environmental	Designed	Dynamic
Training/experience/skills	Visibility	Reaction time	Fatigue
Prior performance on similar tasks	Visual display characteristics	Physical size	Vigilance
	Visual-spatial information available	Reachability	Cognitive workload
	Personal protective gear	Physical workload	Physical workload
	Rank within organization	Short term memory	Short term memory
		Rank within organization	Search time for information in display
			Reaction time
			Incentive

2.2 Task Allocation

In multirobot systems, Parker [56] defined three approaches to tackling the problem of task allocation: bioinspired, organizational, and knowledge-based. The approach taken in the reported research is based on organizational and knowledge-based approaches.

In *bioinspired* approaches, observations made on animal/insect behaviors are applied to solve the problem of task allocation in multirobot systems. A commonly used behavior is from the study of ants; the most popular application of ant behavior is the Ant Colony Optimization (ACO) [24] technique, which was inspired by the foraging behavior of ants. Similarly, some animal/insect behaviors can be applied to the task allocation problem in multirobot systems. Robots in bioinspired approaches are typically homogeneous and exist in large numbers (i.e., swarms). Individually, each robot possesses very limited capabilities. However, when they are grouped together in swarms and interact as a collective, a group-level intelligent behavior emerges. Because it is assumed that every robot has the ability to sense the relevant information in their environment (i.e., *stigmergy* [48]), communication among the robots is reduced significantly. Even in situations when stigmergy is not available, robots only need to broadcast minimal information about their state or environment, thus, there is no need for the robots to communicate about task allocation. A task is allocated when a robot senses that a task needs to be performed and proceeds to perform it. Should a robot fail when performing a task, another robot simply replaces the failed robot. By following this basic behavior, a collective of these robots can achieve the overall system goal. Work such as [6, 43, 44, 48, 49, 57, 76] employ homogeneous robots to solve specific problems. However, re-engineering the solution to a different problem typically involves different types of robots and algorithms. Furthermore, building homogeneous robots that are able to perform a large variety of tasks is more expensive than building different robots that can perform a subset of those tasks.

Organizational approaches use organizational theory for task allocation in multirobot systems and consist of two sub-approaches: role-based and market-based. Role-based approaches employ the use of roles to divide up the work that needs to be done. A role can consists of one or more tasks that need to be completed. Once the set of roles have been defined, robots select (or are assigned) the roles for which they are best suited. Pure role-based approaches typically predefine the set of agents that can perform a particular role; thus, there is no need to determine the set of agents that can perform a particular role at runtime. Role-based approaches that that determine role-agent mappings at runtime typically employ ontology/semantic information. Examples of pure role-based approaches are [70, 75], where [75] assumes homogeneous robots while the [70] assumes no overlap of among roles.

Market-based approaches use principles and theories of market economies to enable robots to negotiate with other robots on which tasks they should perform. Most approaches use a utility function and/or a cost function for computing the approximate values to performing some action. Once the values for a given task have been computed, robots with the highest (utility) or lowest (cost) value are chosen to perform the task. A large majority of market-based approaches [12, 20, 28, 15, 17, 82, 83, 91] share similar concepts as each is broadly equivalent to the previous description of market-based systems; the differences tend to lie in the utility/cost functions, the communication protocols, and the architecture used.

Knowledge-based approaches share ontological and/or semantic information among the robots as the basis for task allocation. Typically, this ontological and/or semantic information have some relation to the system

tasks. Through the process of sharing this ontological and/or semantic information, robots can obtain enough information about other robots to help compute the appropriate robot for a given task. One type of ontological and/or semantic information that is typically shared is robot capability. By sharing knowledge of robot capabilities, knowledge-based approaches are able to include heterogeneous robots when allocating tasks. Many knowledge-based approaches [26, 55, 77, 89, 82, 84, 45, 79, 16] are similar to the approach taken in this research. For example, individual agent's capabilities are captured and communicated to other agents.

The work in this research project was based on the Organizational Model for Adaptive Complex Systems (OMACS) [19], which captures the knowledge required to allow a team of autonomous agents to adapt to failures or changing goals in the form of an organizational model. As shown in Figure 1, an OMACS *organization* consists of *goals*, *roles*, *agents*, and *capabilities*. *Goals* are high-level descriptions of what the system is supposed to accomplish [57]. *Roles* are high-level specifications on how to achieve specific goals. *Agents* are autonomous entities that can perceive and act within their environment [57]. *Capabilities* represent the notion of an agent's ability to perceive and act on its environment.

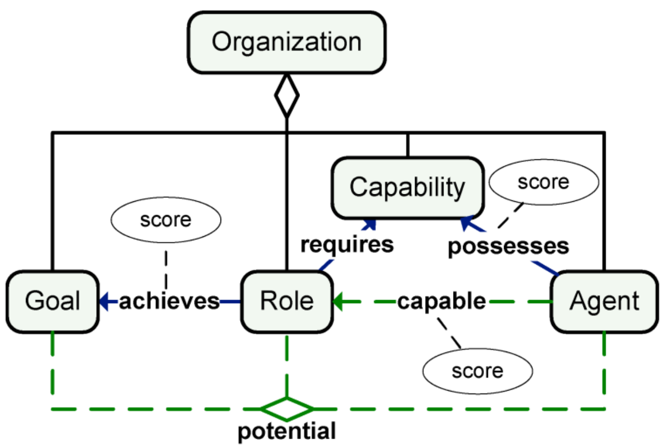


Figure 1. OMACS model

These entities are related via a set of functions: achieves, requires, possesses, capable, and potential. The *achieves* function defines the effectiveness [0.0, 1.0] of a role in achieving a goal, where 0.0 means that the role is unable to achieve the goal. The *requires* function defines the capabilities that a role needs in order for agents to carry out the role's behavior. The *possesses* function defines the effectiveness [0.0, 1.0] of an agent's capabilities, where 0.0 means that the capability is broken or non-existent. The capable function specifies how well [0.0, 1.0] an agent can perform a role, where 0.0 means that the agent is unable to perform the role. The potential function defines how well [0.0, 1.0] an agent can perform a role to achieve a goal, where 0.0 means that the agent is unable to perform the role to achieve the goal. A user definable function (*rcf*), computes a score [0.0, 1.0] that indicates how well an agent performs a role. OMACS-based systems adapt to failures through the use of the potential function (which uses the *rcf* function) to autonomously make assignments (or to allocate tasks). An assignment is a tuple consisting of one goal, one role, and one agent.

3 Workload Moderator Functions when Directing Humans

The initial research focused on modeling and validating a workload human performance moderator function for a situation in which the human participant was guided (Guided Interaction), or directed through a task by either a human or a robotic partner [36, 37]. The human participant was able to ask questions of the remote human or the co-located robot partner, but did not instruct or guide the human or the robot partner.

The research focused on analyzing workload for a situation in which a Chemical, Biological, Radiological, Nuclear or Explosive (CBRNE) incident had occurred and multiple non-ambulatory victims require triage. This research assumed that the volunteer had little first aid training, no prior experience with robots, and formed an ad-hoc team with either a human first responder located outside of the contaminated area or the co-located robot, depending upon condition.

3.1 Scenario Background

The research scenario was a victim triage task for a mass casualty CBRNE incident. CBRNE incident response procedures dictate that first responders are not to enter the contaminated incident scene until a decontamination site is established, the potential hazards are identified, and personal protective equipment is donned [53]. Any response delay can result in significant civilian injury or death. Robots have the potential to enter the scene immediately in order to provide immediate feedback regarding victim locations, triage victims, etc. Such information can assist human responders in determining the appropriate response, including locating, treating, and transporting victims [40]. While robotic technology is not yet capable of all these tasks, robots may be able to enter a scene, identify an uninjured ambulatory victim, and recruit that victim to assist with victim triage. Such human volunteers can assist with locating and triaging victims. The robot can relay information from the scene to first responders located outside the contaminated area.

The START [7] triage system is commonly employed to triage victims during emergency incidents [23, 71]. The START steps require about 60 seconds when completed by a trained responder and focus on assessing the immediacy of care required for a particular victim. START involves a number of steps intended to classify a victim into one of four triage levels: Minor, Delayed, Immediate and Expectant. Minor implies that the victim is ambulatory and coherently responsive. Delayed victims can survive, while waiting up to a few hours for care. Immediate indicates that the victim needs to be treated as soon as possible. Expectant specifies that the victim has passed away or will soon expire. The first step determines if the victim is breathing or not (Respirations). If the victim is breathing, the number of breaths per minute is measured. If the victim is not breathing, an attempt is made to open the airway. A non-breathing victim is classified as Expectant and a breathing victim with over 30 breaths per minute is classified as Immediate. A blanch test or the victim's pulse is measured for all other victims. A blanch test requires the responder to press the victim's fingernail until the color fades, let go and measure the time until normal color returns. If the fingernail takes longer than two seconds to refill, then the victim requires immediate care. Alternatively, the victim's pulse can be measured. If the pulse is not present or is irregular, the victim is classified as Immediate. If the victim remains unclassified, the first responder assesses the victim's mental responsiveness by asking a question or

asking the victim to open and shut the eyes. If the victim is unresponsive, he or she is classified as Immediate. If the victim is responsive, the classification is Delayed.

3.2 The Workload Moderator Function Model

The two conditions modeled in IMPRINT Pro represent a team-based scenario involving first responders (both robot and human) instructing an ambulatory, uninjured victim who is located in the contaminated incident area to perform triage on nearby non-ambulatory victims. The models are specific to two member teams, a human-human (H-H) condition and a human-robot (H-R) condition. The models represent the task activities and the uninjured volunteer's workload. IMPRINT Pro's workload HPMF is divided into seven channels (Cognitive, Auditory, Visual, Fine Motor, Gross Motor, Speech, Tactile), with assignment guidelines provided by the IMPRINT Pro documentation [41].

The Guided Interaction triage scenario requires the uninjured victim to perform the START triage steps on six victims with differing levels of required triage in Round 1 and repeat the triage steps on all victims who were not classified as Expired during the initial triage (five victims) in Round 2. During the second triage round, the order of attending to victims is based on triage level, with those having the most severe triage classification being visited first. Each of the six victims had an assigned triage level and participants were instructed as to the required steps to determine the triage level. The triage levels of two victims may be the same, but the sequence of steps required to reach the decision varied depending on the initial breathing and pulse rate test results. The modeled H-H scenario assumed that the uninjured victim had contacted 911 to report the incident and had volunteered to assist a remote (e.g., located outside of the contaminated incident area) first responder with the triage task. The scenario further assumed that the uninjured victim communicated with the remote first responder via cell phone and that the remote first responder provided step-by-step instructions that led the uninjured victim through the triage steps. The uninjured victim provided responses that were recorded by the remote first responder to assist with incident response planning.

The modeled H-R scenario assumed a robot deployed into a contaminated incident area had discovered the uninjured human victim, who volunteered to assist the robot with the triage task. The uninjured victim executed the instructions provided by the robot and reported results to the robot. The robot reported this information, as well as the location of the injured victim to remote first responders. The robot communicated with the uninjured responder using voice interaction.

Both scenarios used the same task, which was to perform an initial triage assessment on and classification of the injured victims before conducting a follow-up triage. The victim order, provided triage instructions, and victim information were identical across the conditions, except that the H-R model took into account the robot's slower speech pace and an extra step of placing a triage card on each victim with a color representing the triage level.

The IMPRINT Pro models iterate through each atomic task for the entire Guided Interaction Team scenario. Tasks included each START triage step for the individual victim's needs, broken into discrete, atomic tasks based on each individual action the volunteer took. For example, participants counted the number of breaths a victim took in one minute, which was decomposed into discrete, atomic tasks. Participants were instructed to watch the victim's chest rise and fall for one minute, while counting the number of breaths and listening for the teammate to say "stop." When stopped, the participant reported the total number of breaths counted.

For each atomic step, the modeler specifies a running time, title and workload values. Activities such as speech, walking, reading and listening use IMPRINT Pro's built in calculator to estimate how long an average person takes to say a specific number of words, walk a certain number of feet, etc. These times are based on empirical data sets. Each atomic task associated with the uninjured victim had an associated workload value for each of the seven channels: Cognitive, Auditory, Visual, Fine Motor, Gross Motor, Tactile and Speech. A numeric value was assigned to each channel, which resulted in an overall workload associated with a particular atomic task. Each channel had an independent value scale and predefined guidelines for choosing an associated value. For example, the cognitive channel scale is 0 (minimum) to 7 (maximum) cognitive workload. Each channel had an independent value scale and predefined guidelines for choosing an associated value. IMPRINT Pro provides guidelines for assigning these values. When the model executes, assigned workload values for each task are in effect during the entire execution for each atomic task.

The scenario focused on the results for the six injured victim triage levels in two rounds – 11 time periods of measurement. Once the model completed execution, the model output the list of tasks completed by the uninjured victim when triaging the injured victims. Along with each atomic task, the results included the time required to complete the task and the associated workload value for each workload channel and an overall workload value. Figure 2 provides the total time taken to complete triage tasks for each victim in the H-H and H-R models.

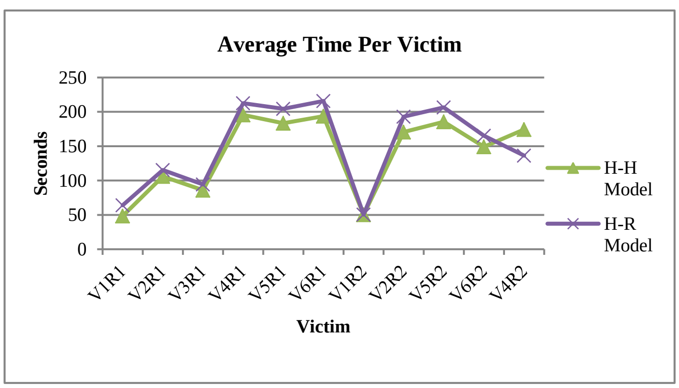


Figure 2. Predicted triage time by victim for each model.

The model output provides a graph of the total workload over the entire scenario and a breakdown of each action at each moment for every human and robot task modeled. Model workload for each victim assessment

was calculated via a weighted average of the workload by the amount of time spent on each atomic task. Figure 3 provides the total workload output for the H-H team, while Figure 4 provides the total workload for the H-R team, both are displayed using the same timescale. During each victim assessment, the workload changes correspond to the individual task demands and the black boxes indicate the time periods when the volunteer triaged the victim labeled at the top of the box (e.g., V1R1). The number after the "V" indicates the victim assessed and the number after the "R" indicates the round, for example V1R1 represents Victim 1 during Round 1.

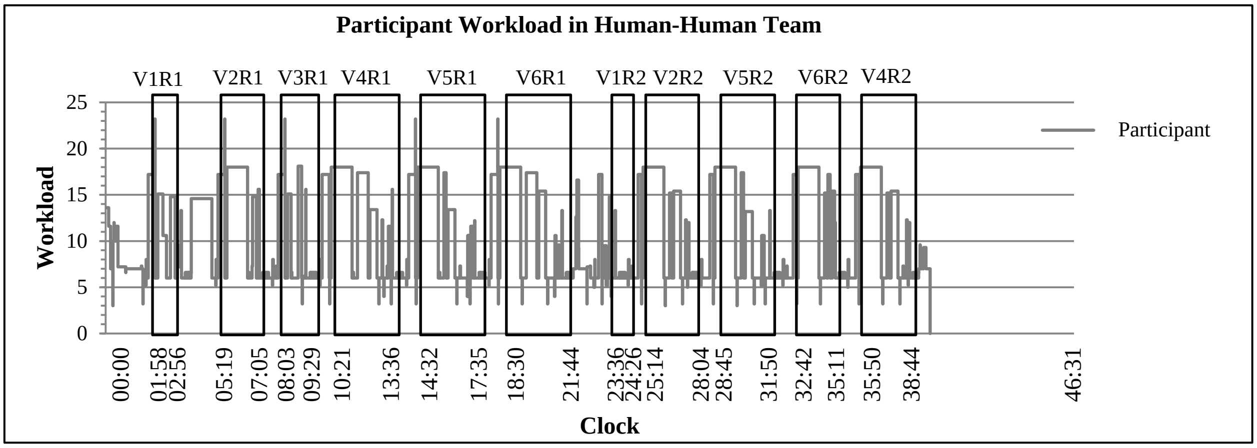


Figure 3. The model output for the H-H Guided Interaction Team model workload for each task and victim.

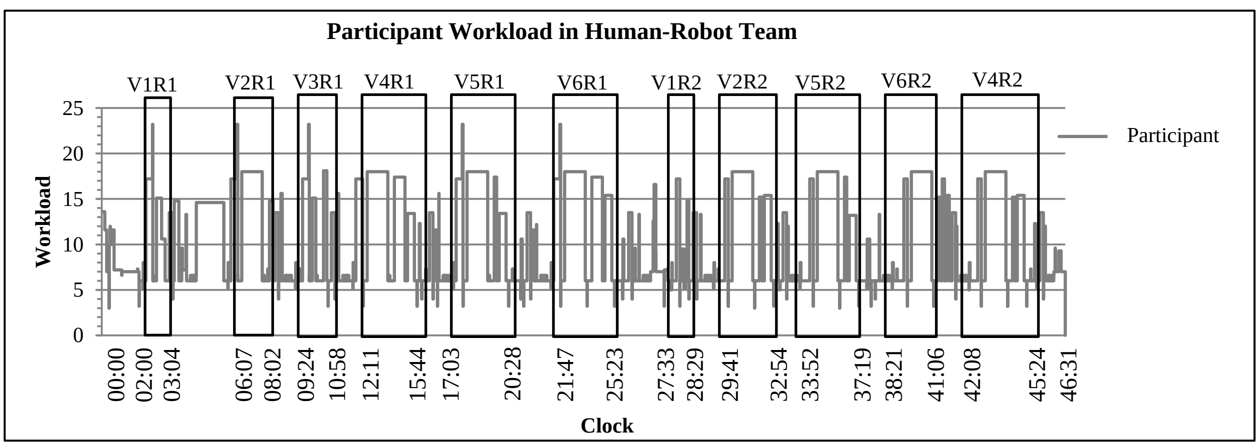


Figure 4. The model output for the H-R Guided Interaction Team model workload for each task and victim.

It can be seen in Figure 5 that overall the H-R team was predicted to require more time to complete all tasks. Both teams experienced very similar trends in total workload with peaks and valleys in similar spots within each victim assessment period; however, the total modeled workload differed due to the time weight for each task. Figure 5 provides the total workload score for each victim for both models. Figure 6 presents the workload values from each individual workload channel for both the H-H (sub-figure a) and the H-R models (sub-figure b).

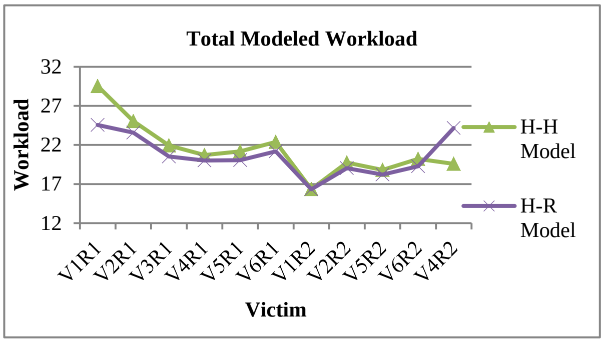
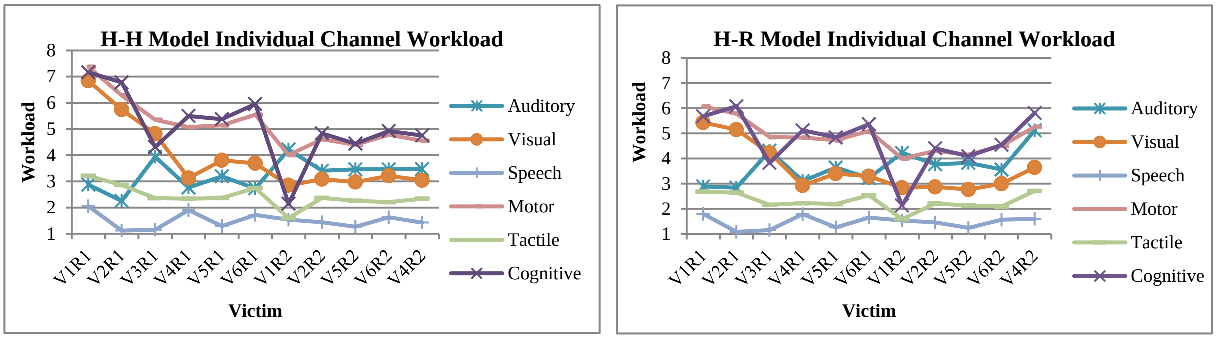


Figure 5. Total workload from the H-H and H-R models.



(a) (b)

Figure 6. Individual workload channel values for the H-H (a) and the H-R models (b).

The models predicted changing workload values based on the specific victim assessed. The models also predicted that total workload for both conditions followed the same general trend independent of teaming partner, either human or robot. The H-R team was expected to exhibit lower overall workload based on the modeling results (Figure 5).

3.3 Validation Evaluation Apparatus

A validation of the modeled workload results was conducted in order to compare the workload metrics from human user evaluations for both conditions, H-H and H-R. The model development and evaluation *hypothesis* was that a measurable difference in workload across the validation conditions exists. The purpose of the evaluation was intended to understand the workload differences between the two conditions and to understand how well the workload HPMF IMPRINT Pro model predicted actual human workload. The evaluation required an uninjured victim to form an ad-hoc team with either a remote human first responder, in the H-H condition or with a co-located mobile robot, in the

H-R condition. The uninjured human received triage instructions from and provided necessary responses to the teammate while completing the triage steps.

3.3.1 Experimental Design

The experimental design was a mixed design with the participants as a random element. The experimental condition (H-H vs. H-R) differed between-subjects and the within-subject element included the series of triage victim assessment tasks. The independent variables were the experimental condition, the victim triage levels, the triage round (i.e., either the initial triage or the follow-up triage) and participant age, gender, experience with robots and first aid training. The dependent variables included both subjective and objective measures (see Section 3.3.3 for details). The evaluation conditions corresponded to the two models (see Section 3.2). The H-H condition participants were completed prior to the H-R condition participants. During the H-H condition, an evaluator played the role of a first responder located outside of the contaminated area. The evaluator provided instructions to the uninjured victim – the participant. The H-R condition paired the participant with a robot. Both the participant and the robot were co-located in the contaminated incident area. A remotely-located human evaluator supervised both the participant and the robot.

3.3.2 Participants

Twenty-eight participants completed the evaluation, fourteen in each condition. All participants had at least some college education and were recruited by flyers around the Vanderbilt University area. Participant compensation was $15 for the approximately 90 minute evaluation. The participants were nearly evenly split by gender across the two conditions, with six males and eight females in the H-H condition and eight male and six females completing the H-R condition. The average age of all participants was 25.2 and age ranged between 18 and 57 years. The H-H condition mean age was 24.2 years and the H-R condition mean was 26.2. The participants rated their level of first aid experience on a Likert scale, with 1 representing no experience and 9 representing an expert level of experience. The average level of first aid experience was 3.6, with the H-H condition mean = 3.3 and the H-R condition mean = 3.9. All participants rated their level of robotics experience on the same scale. The average experience level was 2.7, with the H-H condition mean = 2.9 and the H-R condition mean = 2.6.

3.3.3 Evaluation Metrics

Existing results indicate that heart rate variability (HRV), heart rate, and respiration rate can be employed to assess workload [1, 59, 81], with HRV cited as a reliable measure of workload [61]. While no physiological measure perfectly reflects changes in workload, correlations between HRV and subjective workload measures have proven significant. Secondary tasks and subjective workload measures have also proven useful for assessing workload [31].

The objective metrics included: physiological data from a portable BioHarness ECG monitor [7] (HRV, breathing rate, beat-to-beat interval, heart rate, respiration rate, skin temperature, posture, vector magnitude data and acceleration data), time spent assessing each victim, correctness of

responses to the secondary task questions, and accuracy of triage assessments. Subjective metrics included workload ratings collected after triaging each victim, post-experimental questionnaire responses and post-experimental NASA-TLX [33, 51] responses. Only the HRV, heart rate, respiration rate, secondary task questions, in-task workload rating questions and NASA-TLX responses are reported, please see the evaluation technical report for full results [36].

The secondary recognition task questions were based on a list of five names participants were asked to memorize during the pre-trial briefing: Kathy Johnson, Mark Thompson, Bill Allen, Tammy Hudson and Matt Smith. The names represented a hypothetical team that the participant needed to meet with for debriefing. Participants were given one minute to memorize the list before viewing a briefing video, and another minute to study the list after the briefing video. Thirteen questions incorporating names from the list and other names not on the list were posed throughout the trial. An example question is: "Megan Garner is now setting up the medical treatment site. Was she on the list of names you were given?"

After completing the triage steps for a particular victim, the participants were asked to rank six workload channels on a scale from 1 (little to no demand) to 5 (extreme demand). The six workload channels were Cognitive, Auditory, Visual, Tactile, Motor and Speech. Each channel was defined during the first set of questions. The questions were adapted from the Multiple Resources Questionnaire [10] and the channels were chosen to facilitate comparison to IMPRINT Pro's seven workload channels. In order to prevent participant confusion, Imprint Pro's fine and gross motor channels were combined into the Motor channel rating. When the results were compared to the IMPRINT Pro predicted values, the fine and gross motor channels were added together.

Each IMPRINT Pro workload channel has a specified scale, starting from 0, as minimum workload and ranging from 4 to 7 as the highest workload. The first step in converting the two measures of workload (from the IMPRINT Pro model and the subjective results) to a unified scale added 1 to all IMPRINT Pro workload values. Depending on the specific channel's scale, the subjective workload ratings for that scale were multiplied by a fraction with the highest number on the IMPRINT Pro channel's scale on top and 5 on the bottom. Thus, the workload values are re-scaled for easy comparison. The conversion process for the IMPRINT Pro ratings is shown below where: CMR stands for Comparable Model Rating and IWR stands for a victim's overall IMPRINT Pro workload rating.

$$CMR = IWR + 1 \tag{1}$$

The conversion equation for the in-task subjective workload ratings is shown below. CSR stands for Comparable Subjective Rating and HCV stands for the highest channel value possible in IMPRINT Pro. Both equations are performed for each individual workload channel and the total workload ratings.

$$CSR = Rating * (HCV / 5) \quad (2)$$

The NASA-TLX questionnaire was completed at the end of the entire evaluation.

The time spent triaging each victim was determined by recording a start time for each victim as soon as the participant indicated that he or she had reached a victim. The ending time was determined by noting when the participant finished the last triage task, but before the responder (human or robot) announced the triage level of the victim.

3.3.4 Experimental Environment

The evaluation occurred in the Center for Experiential Learning and Assessment within Vanderbilt University's School of Medicine. During the evaluation, the lights were dimmed and a background noise track created a more realistic environment. The background track incorporated explosion noises, street noise, people coughing and screaming, construction noise, and sirens. The volume was low enough that participants were able to clearly hear the teammate.

Six medical mannequins with various mechanical capabilities were distributed in the evaluation room. Table 2 details each mannequin's capabilities. Figure 7 shows the four of the six mannequins, each identified by the corresponding victim number.

Table 2. Capabilities of each mannequin.

Mannequin	Age	Gender	Capabilities
Victim 1	Newborn	Female	Crying
Victim 2	Adult	Female	Pulse; Respiration Rate; Voice
Victim 3	Child	Male	None
Victim 4	Adult	Male	Pulse; Respiration Rate; Voice
Victim 5	Toddler	Female	Pulse; Respiration Rate
Victim 6	Adult	Male	Pulse; Respiration Rate; Blinking Eyes

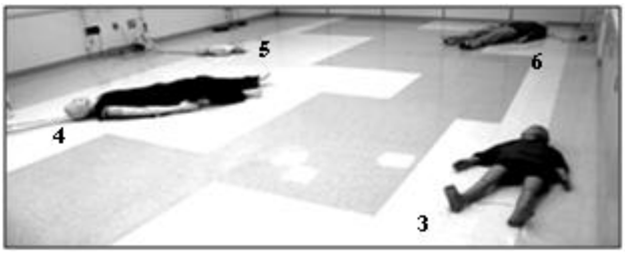


Figure 7. The evaluation environment with four of the six medical mannequins.

The Pioneer 3-DX robot teammate was equipped with a laser range finder and navigated the room autonomously on a pre-planned path [50, 64]. The robot's speech was scripted and controlled by the evaluator [64]. When the robot asked a question, the evaluator logged the response in the robot's script and moved the speech process to the next instruction.

3.3.5 Method

After completing initial forms and questionnaires, participants donned a BioHarness ECG monitor [7]. A baseline heart rate was measured, after which all heart rate channels were recorded continuously

throughout the evaluation. Once the heart rate monitor was functioning properly, a script was read that introduced the disaster response scenario and informed the participant that he or she would be working with a teammate (either human or robot).

Each participant viewed a four minute video, setting the scene of a mass-casualty incident. The video showed scenes from David Vogler's live footage from the September 11th attacks in New York City [85]. After the video, the participant was instructed that his or her role was an uninjured, ambulatory and "contaminated" victim that is unable to leave the contaminated incident area until responders set up a decontamination area.

During the briefing, the participants assigned to the H-H condition were told that they had called 911, could not yet leave the contaminated area, that human responders were not permitted into the contaminated incident area, and asked if they would be willing to assist a human first responder to triage victims. Participants were told that they were transferred to a human first responder who would lead them through the triage steps, and record the participant's responses to questions and the GPS location of the victims based on the participant's cell phone GPS signal. The participants identified which victim to treat next. The participants used a walkie-talkie with a headset and microphone (in place of a cell phone) to communicate with the remote human teammate - an evaluator acting as a first responder. The evaluator was in a remote location, from which she could not be seen or heard.

During the H-R condition briefing, participants were told they would be co-located with the robot because human responders were not permitted in the contaminated incident area. They were asked if they would be willing to work with the robot. The robot led the participants to the victims. The robot communicated with the participants using a digitally synthesized voice projected by a speaker mounted on the robot, while leading the participants through the tasks. The robot's speech was monitored and advanced by the remote evaluator. The participants wore a wireless microphone that transmitted responses to the voice interaction system and the evaluator. The participants were able to ask questions and in a Wizard of Oz manner, the remote evaluator either had the system repeat the robot's statement/question or provided a pre-programmed response.

The victims were positioned such that it almost forced the H-H condition participants to visit the victims in the same order as the H-R condition during the initial triage, see Figure 7. It was possible for participants to visit victims in a different order than planned during the H-H condition. If this occurred, usually a switch of Victims 3 and 4, the evaluator adjusted the script to assess the alternate order during the first round. During the follow-up triage (Round 2), the first responder provided instructions to the H-H condition participants that guided them to the proper victim based upon the initial triage results and the GPS location collected from the participant's "cell phone."

The triage instructions provided and questions asked were identical across conditions. The teammate guided the participant through the steps to identify a victim's triage level. The participants in both conditions started at the same position in the room and moved from victim to victim during the initial triage (Round 1). After completing the initial triage of all six victims, the participant was led back to

the five surviving victims for a second triage check (Round 2). During the second triage for the H-H condition, the next victim was specified by referring to the victim by the order in which they were first visited, for example, “please go to the first victim you triaged.” The robot led the participant to the appropriate victim during the H-R condition. Upon reaching a victim, the teammate provided a summary and led the participant through the triage assessment again. The H-R condition required the participant to place a color-coded triage card on the victim upon completing the triage. The cards were located on the robot platform and the robot instructed the participant which color card to choose. The H-H condition participants were simply told the victim’s triage level. Table 3 details the mannequin settings for each victim, their age, expected triage level, the order each victim was visited, and the type of symptoms for each victim by triage round. Respiration rate is represented as breaths per minute (bpm). Note that the victim triage order during the second triage round was ordered by the most severe triage level after the initial triage.

Table 3. Victim settings for each round, in order visited.

Round	Victim	Triage Level	Details
1	1- Newborn	Immediate	Cries when mouth is opened
	2 - Adult	Immediate	Breathing at 40 bpm
	3 – Child	Expectant	Not Breathing
	4 – Adult	Delayed	Breathing at 20 bpm; Regular Pulse; Responsive
	5 – Toddler	Immediate	Breathing at 18 bpm; Regular pulse; Not responsive
	6 – Adult	Delayed	Breathing at 28 bpm; Regular pulse; Responsive
2	1- Newborn	Expectant	Not breathing; Not responsive
	2 – Adult	Delayed	Breathing at 19 bpm; Responsive
	5 – Toddler	Immediate	Breathing at 18 bpm; Regular Pulse; Not Responsive
	6 – Adult	Immediate	Breathing at 28 bpm; No pulse
	4 – Adult	Immediate	Breathing at 11 bpm; Not responsive

All victims were triaged during Round 1. The third victim was not triaged during Round 2 because this victim was classified as Expectant during the initial triage. Four of the five remaining victims’ triage levels changed prior to the second triage. The secondary task required the participants to memorize a list of names (see Section 3.3.3). Throughout all triage tasks, the participants were asked a question related to the list of names.

After triaging a victim, participants responded to the subjective workload questions. After completing the questions, the participant proceeded to the next victim. Upon evaluation completion, a post-experimental questionnaire and the NASA-TLX workload questionnaire were completed. A second baseline heart rate was collected and the heart rate monitor was removed after questionnaire completion.

3.4 Validation Evaluation Results

3.4.1 Physiological Results

The HRV was analyzed by triage level and condition. The overall H-H mean was 122681.1 (St. Dev. = 424825.9) ms^2 and the H-R mean was 264364.5 (St. Dev. = 875409.6) ms^2. The data was not normally distributed, as evidenced by a Shapiro-Wilks test. Non-parametric Kruskal Wallis tests were

performed to investigate the main effect of Triage Level and condition. The mean HRV by triage assessment are plotted in Figure 8, error bars represent one standard deviation above and below the mean. The x-axis represents the victim assessed and is abbreviated by the victim's number and round number. Mean (M.) and standard deviation (St. Dev.) for HRV, heart rate and respiration rate by triage level and condition are provided in Table 4.

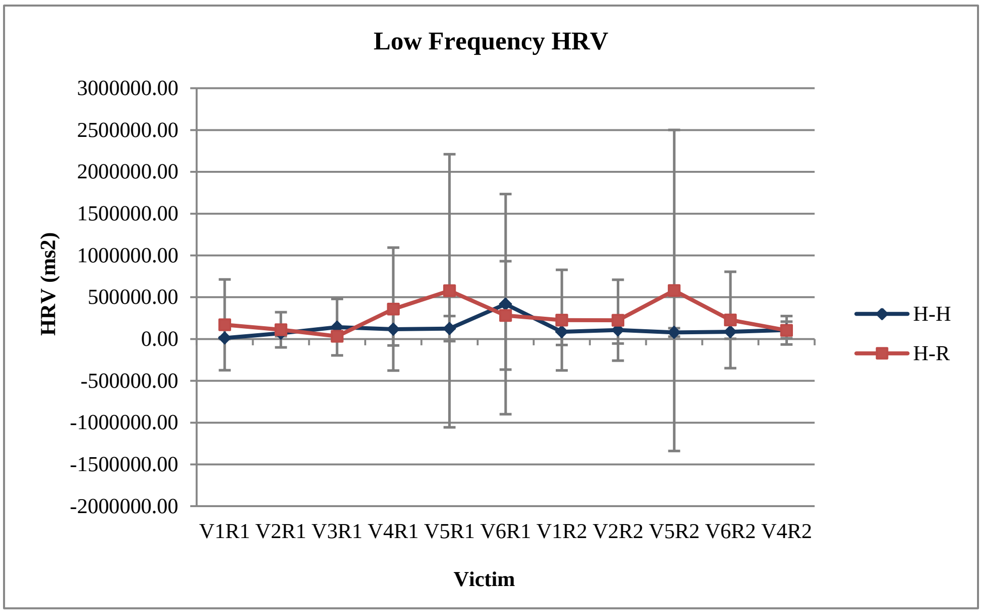


Figure 8. Mean HRV by victim and condition.

Kruskal Wallis tests assessed the main effects and interaction of both condition and triage level, with HRV as the dependent variable and both condition and triage level as independent variables. Results showed a significant main effect of triage level on HRV, $\chi^2(2) = 6.93$, $p = 0.03$. There was no main effect of condition on HRV or interaction effect of triage level and condition. A set of pairwise Wilcoxon post-hoc test indicated that Delayed victims elicited higher HRV than Expectant ($p = 0.018$) victims. There were no other significant comparisons.

The heart rate descriptive statistics by condition and triage level are provided in Table 4. A t-test found that the H-H condition had a significantly higher heart rate, $t(306) = 3.59$, $p < 0.01$. A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with heart rate as the dependent variable and both condition and triage level as independent variables. Results showed that H-H heart rate was significantly higher than that of the H-R condition, with $F(302,1) = 12.78$, $p < 0.01$. There was no main effect of triage level and no interaction effect of triage level and condition.

The respiration rate descriptive statistics separated by condition and triage level are also provided in Table 4. A t-test found that the H-H condition had a significantly higher mean respiration rate, $t(306) = 2.65$, $p = 0.01$. A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with respiration rate as the dependent variable and both condition and triage level as

independent variables. Results showed that the H-H respiration rate was significantly higher than H-R, $F(302, 1) = 6.93$, $p < 0.01$. No main effect of triage level on respiration rate or interaction effect of triage level and condition were found.

Table 4. Descriptive Statistics for all physiological metrics. St. Dev. are provided in parentheses.

Metric	Triage Level	H-H	H-R
HRV (ms^2)	Delayed	217111.9 (777135.7)	288511.1 (618166.9)
	Immediate	78250.5 (88862.4)	297568.5 (1077475.2)
	Expectant	115992.5 (263014.1)	129718.6 (43011.9)
	Overall	122681.1 (424925.9)	264364.5 (875409.6)
Heart Rate (beats per minute)	Delayed	85.03 (12.21)	81.13 (9.40)
	Immediate	85.20 (12.00)	80.42 (10.27)
	Expectant	87.36 (12.65)	82.31 (11.45)
	Overall	85.55 (12.13)	80.96 (10.22)
Respiration Rate (breaths per minute)	Delayed	18.88 (3.69)	18.01 (2.64)
	Immediate	19.13 (3.91)	17.71 (3.25)
	Expectant	19.01 (4.50)	18.69 (3.25)
	Overall	19.04 (3.94)	17.97 (3.10)

3.4.2 In-Task Subjective Workload

The individual channel subjective workload ratings gathered at the completion of each victim assessment were combined into a total workload value. The overall mean for H-H workload was 16.26 (St. Dev. = 6.31), while the H-R workload mean was 13.48 (St. Dev. = 4.80). A t-test comparing the overall results across conditions found no significant difference. The mean H-H condition workload for Delayed victims was 19.31 (St. Dev = 5.94) and 15.02 (St. Dev. = 4.70) for the H-R condition. The mean for the H-H Immediate victims, was 16.06 (St. Dev. = 5.89) and for H-R was 13.48 (St. Dev. = 4.60). The Expectant victims H-H mean was 15.21 (St. Dev. = 5.97) and the H-R mean was 11.14 (St. Dev. = 4.74). Figure 9 provides the total workload for each condition at each victim assessment point.

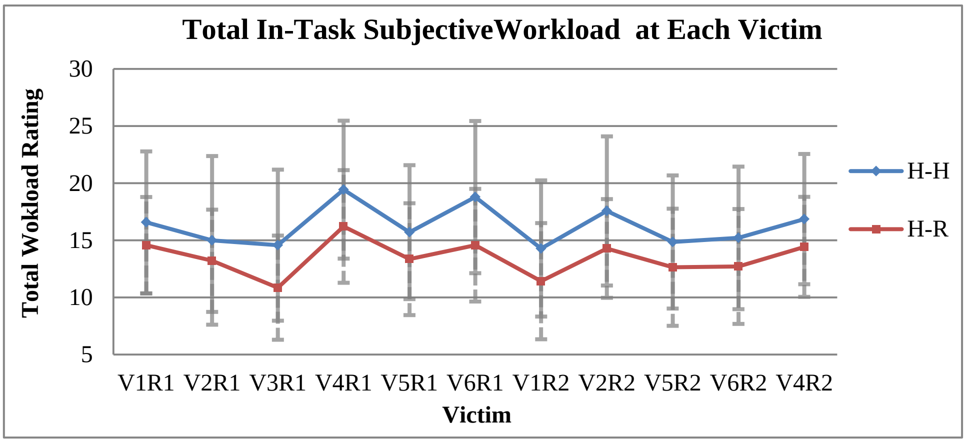


Figure 9. Total workload by victim and condition.

A two-way ANOVA assessed the main effects and interaction of both condition and triage level, with the total in-task subjective workload ratings as the dependent variable and both condition and triage

level as independent variables. Results showed a significant main effect of triage level on workload ratings, with $F(302,2) = 10.29$, $p < 0.01$. There was a main effect of condition on the workload ratings with $F(302,1) = 29.86$, $p < 0.01$, showing that the H-H workload ratings were significantly higher than the H-R workload ratings. There was no interaction effect of triage level and condition. A Tukey HSD test showed that the significant difference between triage levels was due to the Delayed victims being rated significantly higher than both the Immediate ($p < 0.01$) and Expectant ($p < 0.01$) victims. There was no significant difference between the Expectant and Immediate workload ratings.

3.4.3 Secondary Task Questions

The number of correct answers to secondary task questions was compared between conditions. Thirteen questions (Q.) were asked in total – one during the introduction to the task (Q. 1), one during the triage of each the victim during Round 1 (Q. 2–7), one between the two rounds (Q. 8), and one during the triage of each victim during the second round (Q. 9-13). Overall, the mean number of correct responses was 12.71 (St. Dev. = 0.61) during the H-H condition and 12.43 (St. Dev. = 0.65) for the H-R condition. T-tests across conditions found no significant difference. An analysis conducted based upon triage level found no significant results. The division of correct answers by condition is provided in Table 5.

Table 5. Average number of correct responses by triage level

Triage Level	H-H	H-R
Delayed - Q. 5, 7, 10	2.93 (0.27)	2.86 (0.36)
Immediate - Q. 2, 3, 6, 11, 12, 13	5.86 (0.36)	5.76 (0.43)
Expectant - Q. 4, 9	1.93 (0.27)	1.93 (0.27)
Overall	12.71 (0.61)	12.43 (0.65)

3.4.4 NASA TLX

Each participant completed the NASA-TLX questionnaire. The mean overall weighted score for the H-H condition was 57.38 (St. Dev. = 14.00), while the mean for the H-R condition was 48.59 (St. Dev. = 11.98). A t-test found no significant difference between the overall scores. While this result is not significant, it indicates a trend that those in the H-H condition tended to rate their overall workload values slightly higher than the H-R condition participants.

3.4.5 Correlations

A partial Pearson's correlation was performed to analyze the correlation between HRV, heart rate, respiration rate and the total in-task subjective workload rating from each victim while adjusting for the independent variables of victim being assessed and victim triage level. Across both conditions, in-task subjective workload ratings were significantly negatively correlated to respiration rate, $r(290) = -0.15$, $p = 0.01$ and had a significant positive correlation to heart rate, $r(290) = 0.16$, $p < 0.01$. The correlation between HRV and subjective workload ratings was not significant, $r(291) = 0.102$, $p = 0.08$. The literature [1, 59, 84] implies that these three physiological measures may be able to represent workload. Since the physiological metrics were correlated to the in-task subjective workload ratings, the trends shown by these three physiological measures can be considered when assessing the

difference in workload between conditions. The literature reports a positive correlation between both HRV and heart rate and workload, and a negative correlation between respiration rate and workload.

3.5 Validation Evaluation Results Compared to Model Results

3.5.1 In-task Total Subjective Workload

Total workload was compared between the model results and the validation results.

Figure 10 provides a comparison of the H-H In-task total subjective data and H-H model. The calculated difference between the modeled workload values and the in-task subjective workload values demonstrate how effective the models were at predicting human behavior. The mean delta between the H-H subjective values and the model results at each time point was 0.88 (St. Dev. = 3.84). The mean absolute value of the difference between the model and actual data was 3.23 (St. Dev. = 2.03).

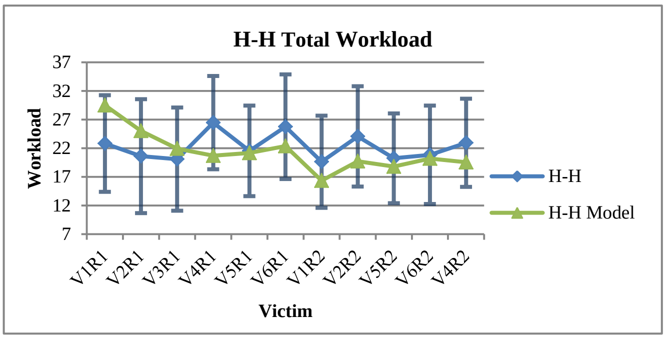


Figure 10. Total rescaled workload ratings for the H-H evaluation, versus the H-H model's total workload value. Error bars represent one standard deviation above and below the mean.

A comparison of the H-R validation In-task subjective total workload and the IMPRINT Pro workload prediction for the H-R model is provided in

Figure 11. The mean delta between the H-R condition and the model was -2.15 (St. Dev = 2.57). The mean absolute value of the difference between the data sets was 2.64 (St. Dev. = 2.01). These results imply that the H-R model was a slightly better predictor of H-R workload than the H-H model.

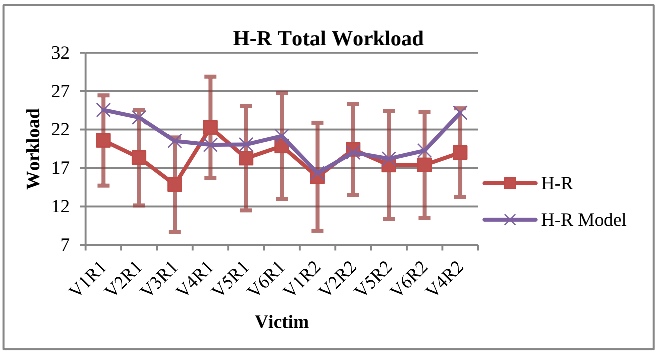


Figure 11. Total rescaled workload ratings for the H-R evaluation versus the H-R model.

3.5.2 Individual Workload Channels

The difference between the model and the empirical results was compared for each individual workload channel. A smaller average difference indicated that the model more closely predicted human performance. The H-R IMPRINT Pro model more closely predicted the actual Auditory, Visual, Speech and Tactile workload data. The two models produced virtually the same accuracy in predicting Cognitive workload.

IMPRINT Pro has two motor channels, fine and gross, that were combined in the evaluation. In order to compare this dual-motor channel to IMPRINT Pro's motor channels, all values were scaled to the fine motor channel scale from IMPRINT Pro and the two IMPRINT motor channels were then added together. The H-H IMPRINT Pro model more closely predicted the actual Motor subjective workload data.

3.5.3 Time Spent Per Victim Measured

The time spent triaging each victim for each condition was compared to the models. Figure 12 compares the H-H evaluation timing data to the H-H model, which is within one standard deviation of the evaluation results for each point except for V1R1 (1.14 seconds (s)), V1R2 (3.53 s) and V5R2 (1.10 s). The mean difference between the H-H evaluation data and the model is -2.32 s (St. Dev. = 18.17). Some model data points are above the evaluation results, therefore, the mean difference when taking the absolute value of the difference at each point is 13.30 s (St. Dev. = 11.89).

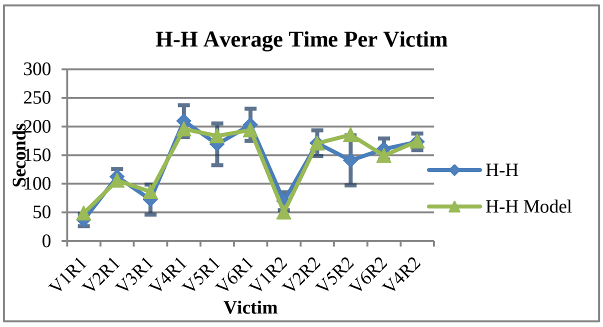


Figure 12. Comparison of average time to triage each victim for the H-H evaluation data and the model.

Figure 13 depicts the H-R timing data gathered from the evaluation compared to the model, which is within one standard deviation of the evaluation data except for V3R1 (38.99 s), V1R2 (16.20 s) and V4R2 (8.08 s). The mean difference between the H-R evaluation data and the model is 0.98 s (St. Dev. = 27.47). Some model data points are above the empirical data, thus the mean difference when taking the absolute value of the difference at each point is 21.95 s (St. Dev. = 15.02).

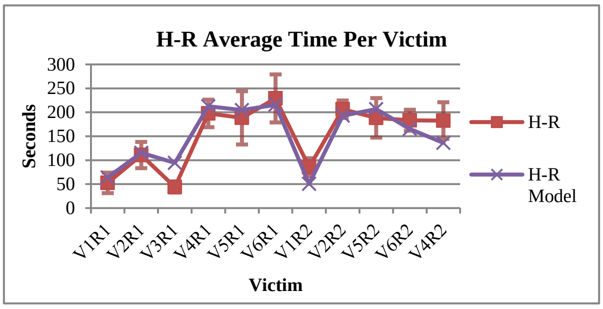


Figure 13. Comparison of average time to assess each victim for the H-R evaluation data and the model.

While the model is not perfect, the predictions generally align with the subjective workload data. When comparing time spent for victim assessment, the model did a good job of providing a close predictor of the required time to assess each victim. The model's predictions are within one standard deviation from the actual mean in eight of eleven cases for the H-H condition and eight of eleven cases for the H-R condition. While the number of data points outside one standard deviation was the same, the H-H model predicted the three points closer to the evaluation data than the H-R model. The H-H model did a better job of predicting the time required to assess each victim.

3.6 Discussion and Findings

The research hypothesis was that a difference between the H-H and H-R condition workload values exists. This hypothesis was partly supported based upon the validation and comparison to the IMPRINT Pro model. The H-R condition resulted in a slightly lower workload level, apparent in the model predictions and the subjective and physiological workload measures.

The predicted IMPRINT Pro model workload trends were very similar to the actual in-task subjective workload rating results. The models employed current workload models and were adjusted for the slight differences in task time between the H-H and H-R scenarios. Overall, the models provided a valuable tool for prediction of workload. The data imply that the H-R condition participants experienced a lower level of workload than participants in the H-H condition, but the models more accurately predicted the H-R subjective workload. Time spent assessing each victim was closely predicted by the models, with the H-R condition overall requiring more time. Current workload HPMFs may be applicable to human-robot peer based teams with the understanding that H-R teams may experience slightly less workload than H-H teams; however, additional analysis of more complex relationships and tasks are planned and required.

Both HRV and in-task subjective workload measures indicated a main effect of triage level on workload, showing that the evaluation's controlled manipulation of workload did affect the participants' physiological responses and subjective ratings.

There was a main effect of condition on heart rate, respiration rate and in-task subjective workload ratings. All of these results indicated that H-H workload was higher than H-R. This difference in workload was corroborated by the NASA-TLX results, showing that while not significant, the H-H

workload ratings tended to be higher than the H-R ratings. The models' predictions align with subjective empirical data, showing a lower level of workload in the H-R condition.

There are currently two hypotheses as to why the H-R condition resulted in lower workload. One hypothesis is that the embodiment of the robot may directly result in lowering the human's workload during the H-R condition. The second hypothesis is that the robot's slower movement speed from victim to victim and slower interaction with the participant during the triage tasks may result in a lower workload for the H-R condition participants.

4 Workload and Reaction Time Moderator Functions for Collaborative Teams

The development of successful collaborative human-robot teams is dependent on a number of factors. One such factor is the impact of workload and reaction time on the collaborative team relationship. The development of collaborative human-robot teams must go beyond the situation where the human is directed by a robot in order to support more realistic and uncertain tasks that require true cooperation and joint decision making. The presented research focuses on quantifying workload and reaction time for collaborative teams in which the team members have some assigned responsibilities, but must also make joint decisions. IMPRINT Pro was used to model human-human and human-robot teams, which was validated via a user evaluation.

4.1 Scenario Background

The Collaborative Team scenario involved conducting a reconnaissance of an academic building after a bomb threat was received. In general, upon receiving a credible bomb threat, an area is evacuated and carefully, systematically and thoroughly searched for suspicious items and bomb materials. During the preliminary reconnaissance, teams of two typically search for unusual objects, describing and taking note (e.g., pictures) of strange sounds, smells and anything out of place [11, 47]. The teams check for items in containers (e.g., trash cans), behind and underneath items in the environment (e.g., furniture), and in the building structure (e.g., ceiling) [40]. If a suspicious object is found, it is not disturbed and information regarding its whereabouts and characteristics are reported immediately to personnel, including incident command, who are located a safe distance from the incident area.

4.2 The Function Model

The Collaborative Team models were created using IMPRINT Pro for a team of two humans (human-human) and a human-robot team that incorporated the workload and reaction time HPMFs. The models represented the subtasks required to complete the Collaborative Team scenario and incorporated a level of uncertainty related to exactly what tasks were performed during each run of the model. For example, when answering a question, during some runs the human responded "yes," and completed a follow-up action, while in other runs the human responded "no." Adding alternate scenario paths created a range of workload values. The human-human (H-H) and the human-robot (H-R) models differed only in the timing of some tasks. The robot spoke ~1.3 times slower and traveled

the same distance ~1.5 times slower than the human partner, which were incorporated into the H-R model.

Workload was modeled using the same approach as employed for the Guided Interaction Team models (see Section 3.2). Reaction Time was modeled using IMPRINT Pro's micromodels of behavior. Reaction Time represented the time it took for each participant to react to an out-of-place item once it was in the field of view (See Section 4.3.3). IMPRINT Pro provides values for simple reaction times including binary response, physical matching, name matching or class matching. The IMPRINT Pro micromodel does not include the time taken for the visual system to recognize items. The provided micromodel incorporates both the reaction to an item and the decision to respond. These values did not incorporate all aspects of Reaction Time, namely the recognition by the visual system. Therefore, the modeled Reaction Time was based on a combination of micromodel time values. The time of each Item Reaction Time task was the same. The Reaction Time result provided by IMPRINT Pro represents the sum of all the individual micromodeled reaction times.

4.3 Validation Evaluation Apparatus

The models provided a prediction of the human's workload based upon the workload and reaction time HPMFs. A user evaluation was designed to validate the IMPRINT Pro modeling results. The validation evaluation hypotheses were: 1) the models accurately predict the human's workload in both scenarios, 2) workload will be lower in the human-robot team, and 3) response time is not impacted by team partner.

4.3.1 Experimental Design

The experiment was a mixed design with the participants as a random element. The experimental condition (human-human vs. human-robot teams) differed between-subjects and the within-subject element included a series of Investigation Areas. The evaluation environment was divided into six Investigation Areas, each with a corresponding Investigation Index representative of an expected workload level: Low, Medium or High. Each area was assigned a point value based on the Investigation Area contents and required tasks. One point was assigned for a present non-suspicious item that the participant was trained to look at (e.g., empty trash can). Two points were assigned for an item that needed to be discussed by the team members (e.g., a message on a whiteboard). Three points were assigned for an item found by physically moving objects around (e.g., opening a recycling bin lid). An additional point was added for a joint decision. The two Low Investigation Indexed areas had scores of 10, the Medium Investigation Indexed area scores were 15, and the High Investigation Indexed area each scored 20 points.

The independent variables were the experimental condition, the Investigation Indices and participant age, gender, experience with robots and first response training. The dependent variables include both subjective and objective workload metrics (please see Section 4.3.3). All H-H condition participants completed the evaluation prior to the H-R condition participants. During the H-H condition, an

evaluator played the role of the first responder and human teammate. Verbal interactions between the participant and human experimenter were dictated by a verbal script that the evaluator possessed.

The H-R condition paired the participant with a semi-autonomous Pioneer 3-DX robot. The robot drove at ~1 meter per second and spoke at ~2.2 words per second. The robot teammate was equipped with a laser range finder and navigated the path through the hallway autonomously on a pre-planned path supervised by the remote experimenter. Figure 14.a provides an example of a participant working

a. b.

Figure 14 (a) Participant and robot teammate in the H-R condition, assessing a fake bomb underneath an eyewash station (Item 14). (b) Enlargement of Item 14.

with the robot to assess suspicious items.

Autonomous navigation was accomplished by using a SICK LMS 200 laser range finder for object sensing, and a particle filter for localization. Radio frequency-identification (RFID) tags were used to allow the robot to identify objects in the environment that were pertinent to the evaluation. When a particular RFID tag was sensed by the robot, control software was used to map that tag's identification number to a real world item. Knowledge of this item was communicated to participants through the robot’s speech, created using a text-to-speech program running on the robot (See Section 3.5). The robot’s speech was scripted and controlled by the experimenter, and provided the ability to repeat statements. The experimenter was able to insert customized utterances into the script, if needed. When the robot asked a question, the evaluator logged the response in the robot’s script. The robot spoke using a digital female voice with an American accent, which was chosen to be similar to the human experimenter’s voice during the H-H condition.

The RFID hardware utilized was a GAO RFID 433 MHz Active RFID reader which interfaced with the robot via the robot's RS232 serial port. The GAO RFID reader was used to sense 433 MHz active RFID tags. The Player/Stage software framework was used to control the robot [29] and a new driver was written to interface Player/Stage with the RFID reader.

The robot traveled the same path as the human evaluator and used the same verbal script to communicate with participants. A human evaluator supervised both the participant and the robot remotely.

4.3.2 Participants

Thirty-six participants completed the evaluation, eighteen in each condition. The 19 male and 17 female participants were not experts in first response or robotics and ranged in age between 18 and 56 years old. The mean participant age of the H-H participants was 27.4 with a standard deviation (St. Dev.) of 8.4 years. The mean participant age for the H-R participants was 24.1 (St. Dev. = 4.6).

Participants rated their experience with search and rescue response on a Likert scale from 1 (little or no experience) to 5 (very experienced). The median response was 1 for both conditions. Participants provided a median response of 1 in both conditions when asked to rate their experience with robots on the same Likert scale. Each participant had normal or corrected-to-normal visual acuity.

4.3.3 Evaluation Metrics

The objective metrics captured physiological metrics via a portable BioHarness electrocardiography monitor [7] (heart rate variability, beat-to-beat interval, heart rate, respiration rate, skin temperature, posture, vector magnitude data and acceleration data), walking data from a Garmin footpod pedometer [27] and time spent searching each Investigation Area. Additional objective metrics included the correctness of responses to the secondary task questions (STQ), STQ response time, number of times an STQ was repeated, Item Response Time, Item Find Type, the number of items found by participants, the number of photographs taken by participants, and the number of items recalled post-experiment[2].

The reported results for the physiological data analysis include normalized heart rate, normalized respiration rate, and heart rate variability (HRV). Heart rate and respiration rate were normalized by determining the mean difference between a participant's raw heart/respiration rate during the investigation of each area and the mean of a baseline heart/respiration rate captured during training. The absolute value was not taken, because the directionality from baseline was an important indicator of change in respiration rate. Normalized heart rate is measured in beats per minute, while normalized respiration rate is measured in breaths per minute. HRV was calculated by taking the mean ratio of Low Frequency HRV over the High Frequency HRV.

Total investigation time was measured beginning when the human or robot partner was introduced to the participant and ended once the last set of in situ workload questions was completed. Individual area investigation times were recorded beginning when the previous area's in situ workload questions were completed and ended when the next set of in situ workload questions began.

All timing data was determined using video coding from the head mounted camera worn by participants. Two video coders determined the start and end times for each Total Investigation Time, Area Investigation Time, Item Response Time and Secondary Task Question response time. Workload ratings, item find type, items assessed and secondary task question responses were also determined by

[2] Please note that the following objective metrics are not reported in this report: beat-to-beat interval, skin temperature, posture, vector magnitude data, acceleration data, walking data, number of photographs taken, and number of items recalled post-experiment.

video coding. Inter-coder reliability was calculated using Cohen's Kappa to be .802 with weights to accept matches within 3 seconds.

The secondary task question responses represented the participants' spare mental capacity for recognizing items from a list and recalling specific associated values. The participants were given one minute to memorize a list of six chemicals, each with an associated Danger Level:

- Chlorobenzene, 75
- Acetate, 25
- Naphtha, 100
- Ethylamine, 10
- Pyridine, 5
- Ammonia, 50

Twelve questions related to the memorized list were asked, independent of condition. The standard question structure was: "Was X on the list of chemicals you received?" If the participants responded "Yes," regardless of the ground truth, the structure of the follow-up question was, "What was the danger level of X?" The chemicals (represented by X in the questions), and the order of inquiry was: 1. Ethylamine, 2. Acetate, 3. Acetone, 4. Ammonia , 5. Chlorine, 6. Naphtha, 7. Ethane, 8. Propane, 9. Chlorobenzne, 10. Nicotine, 11. Acrolein, and 12. Pyridine. The dependent variable On List Correct represented whether participants correctly answered if an item was on the list and the Danger Level Correct metric represented whether participants correctly provided the items' danger level, if they claimed it was on the list.

Total On List Response Time was calculated by measuring how long it took participants to respond to questions regarding whether a chemical was on the list. The time was measured starting from the end of the first time the question was asked and ending when an answer was provided. This time measurement included the time for the responder to repeat the question or chemical name.

On List Response Time after Repeats was calculated by measuring how long it took to respond to questions regarding whether a chemical was on the list. This measurement does not include the time taken for the participant to ask for question or chemical name repeats. The measurement began once the responder finished repeating the chemical name for the last time and finished once the participant provided the answer.

Danger Level Response Time represents how long participants took to respond to the chemical's danger level questions, if the participant had claimed the item was on the list. If the participant answered the question before the responder finished asking the question, then a negative response time was determined by measuring the time between when the participant said the danger level and when the responder completed speaking the danger level question. All participant responses that responded with a danger level when asked the on list question were eliminated from the analysis.

The number of repeats represented the total number of times that the participant asked the responder to repeat the chemical's name or the entire question. Cases where the participant asked the responder to confirm his or her pronunciation of the chemical name were considered repeats.

The numbers of items found by participants were determined in two ways: the total number of items found and the total number of additional items reported. A total of twenty-six items were specifically located throughout the environment, please see Section 4.3.4 for additional details.

Item Reaction Time for each item was measured by recording the time on the point-of-view camera when the object came into view and subtracting it from the time when an item assessment began. Item Response Time was only calculated for items participants found, thus it does not include any items pointed out by responders. The assessment began in many ways: an extended visual fixation, a verbal utterance about the item, shining the laser pointer on the item, touching the item or taking a photo of the item. These methods of identifying an item are termed the Item's Find Type. The Find Type was considered only for items found by participants.

Item find type refers to the method participants used to indicate that an item was noticed by the participant. These methods included an extended visual fixation, a verbal utterance about the item, shining the laser pointer on the item, touching the item or taking a photo of the item.

The subjective metrics included in situ workload ratings collected after searching each investigation area, post-experimental questionnaire responses[3] and post-experimental NASA-TLX [33] responses.

The in situ subjective workload metric required the participants to verbally rank six workload channels upon completing the search of each investigation area. The six workload channels were Cognitive, Auditory, Visual, Tactile, Motor and Speech and the ratings were provided on a scale from 1 (little to no demand) to 5 (extreme demand). Each channel was defined during training, but participants were able to ask for any channel definition during any of the workload questions. The questions were adapted from the Multiple Resources Questionnaire [10]. The workload channel definitions provided to the participants were:

- Auditory: recognizing words, tones, mood and emotions through sound.
- Visual: recognizing faces and objects, sustaining visual attention, judging distances, spatial reasoning and reading.
- Speech: instances when you used your own voice.
- Motor: movement and control of face muscles, arms, hands, fingers, legs and feet.
- Tactile: recognition and judgment of shapes using the sense of touch.
- Cognitive: making judgments, estimation, learning, problem-solving, decision-making, reasoning, memorization and recalling from memory.

The NASA-TLX questionnaire was completed at the end of the entire evaluation [33]. Due to an experimenter error, data from eight participants in the H-H condition were not recorded.

[3] The post-experimental questionnaire responses are not presented in this report.

4.3.4 Experimental Environment

The evaluation environment was a single floor of a Vanderbilt University academic building. The hallway was divided into six Investigation Areas, as shown in Figure 15. The teams followed the same path through the hallway and traversed each Investigation Area in numerical order. The participants, independent of condition, began the investigation at the location labeled by the star in Figure 15. The remote evaluator responsible for supervising the robot during the H-R condition sat in the room labeled with the triangle.

Suspicious and non-suspicious items were placed in the hallway by the experimenters. The same suspicious items were placed in the same locations for both conditions. Nineteen items (both suspicious and non-suspicious) were stationed throughout the search areas. There were seven additional items incorporated into the search script, which included hazard placards, laser laboratory warning signs, one fire extinguisher and one piece of lab equipment. The 26 items are listed by area in Table 6, which also identifies the suspicious items. Items that were normally resident in the environment (with the exception of hazard placards and laser signs) are not listed in Table 6. Such items include bulletin boards and white boards with nothing of note posted, recycling bins and trashcans without anything additional placed inside and fire extinguishers without any alteration from normal positions. The location of each item is presented in Figure 15 and is labeled with the item number. The figure also delineates the locations of benign bulletin boards, trash bins, recycling bins and fire extinguishers.

Table 6. The items incorporated into the verbal script. Suspicious items are marked with an X in the "Suspicious?" column.

Investigation Area	Item Number	Item Description	Suspicious?
Area 1	1	Map in trash can	X
	2	Backpack on bench	X
Area 2	3	Soda bottle with suspicious material in recycling bin	X
	4	Math equations written on white board	
	5	Box of bomb-making supplies on windowsill	X
	6	Box of textbooks on floor	
Area 3	7	Hazard placard and laser sign over closed lab door	
	8	Bomb-making instructions in a trash can in the large lab	X
	9	Bomb-making materials on counter in the large lab	X
Area 4	10	Printout discussing C4 on bulletin board	X
	11	Fire extinguisher sitting outside of its case	
	12	Hazard placard next to closed lab door	
	13	Additional hazard placard next to another lab door	
	14	Bomb underneath eyewash station	X
	15	Wires hanging from the ceiling	X
	16	Map on bulletin board	X
Area 5	17	Note in emergency door	X
	18	Laser sign over closed lab door	
	19	Laser sign over another closed lab door with a keypad lock	
	20	Lunchbox under water fountain	
Area 6	21	Box of bomb-making supplies under table in hall	X
	22	Message written on whiteboard: "rendezvous code green"	X
	23	Note on windowsill	X
	24	Computer cables in the small lab	
	25	Cleaning supplies in the small lab	
	26	Unknown machine / experimental equipment in the small lab	X

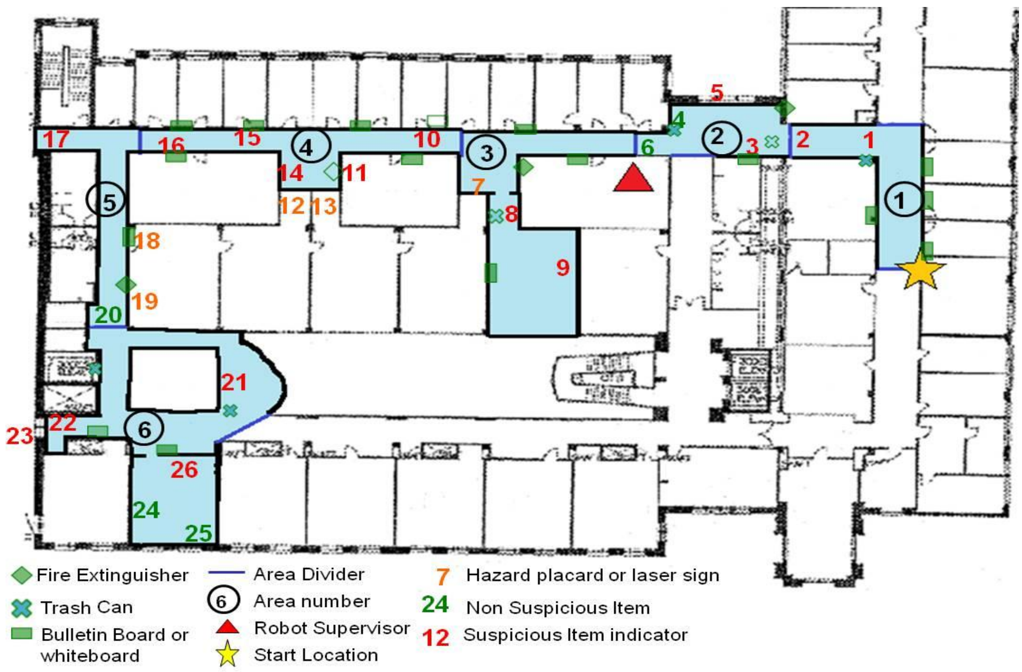


Figure 15. The map of the experimental environment with each of the six Investigation Areas shaded and labeled. The locations of all items in the environment are shown.

Figure 16 shows Item 2, a suspicious backpack placed on a bench in Area 1, which was clearly visible to participants. Some items (e.g., the items placed in trash cans) were not clearly visible and required participants to expend more effort to locate and identify, for example the fake pipe bomb placed under the eye wash station in Figure 14.

Figure 16. A backpack (Item 2) placed on a bench filled with wires.

4.3.5 Method

One experimenter acted as the Trainer, introducing participants to the scenario, helping to don equipment and playing the training video. Upon arrival, participants completed a consent form and demographic questionnaire. After which the participant's height and weight were recorded. The participants were informed that there had been an anonymous bomb threat and that the task was to search for out of place items and report anything suspicious to their partner. The participants donned a BioHarness electrocardiography monitor [7] and a baseline heart rate was measured, after which all heart rate channels were recorded continuously throughout the evaluation. A training video was played indicating how typical searches are executed (e.g., trashcan lids are lifted), what types of items should be deemed suspicious and indicating that photographs should be taken after assessing a suspicious item or hazard sign. The participants were informed that the human or robot partner was responsible for checking fire extinguishers and taking air samples, and that participants were responsible for checking bulletin boards and trash cans, while looking for additional out of place items. The teammates encountered situations requiring joint decision making.

Upon briefing completion, participants donned the remaining equipment, including a neon reflective vest to indicate that the participant was a part of the investigation, a Garmin footpod pedometer and watch [27], a Shure microphone headset, and a Looxcie head-mounted video camera (attached to the microphone headset). The participants were provided with a point-and-shoot digital camera and a laser pointer. The microphone was used to record the participants' speech and during the H-R condition. The participant was instructed to use the laser pointer to indicate what they were investigating (to be recorded by the head-mounted camera). The point-and-shoot camera captured images of potentially suspicious items and hazard placards.

The evaluation began with an introduction speech from either the human responder or the robot partner, after which the team began the search. Periodically the responder partner (either human or robot) informed the participant of air sample readings. It is common to collect air sample readings and

the scenario incorporated the measurement of methane that was found to increase near a research laboratory located in Investigation Area 4.

The responder halted the immediate Investigation Area search when the area boarder was reached. The responder first verified with the participant whether or not any additional items should be investigated after which the participant rated the workload channels (see Section 4.3.3).

Upon completing the investigation, participants returned to the training room (See Figure 15). After removing the vest, pedometer, microphone and camera and returning the point-and-shoot camera and laser pointer, the participants wrote down any item they remembered assessing. Participants completed the NASA-TLX workload questionnaire [33] and the post-trial questionnaire. A second baseline heart rate was collected prior to the removal of the heart rate monitor. Finally, the participant completed a survey regarding expectations of human, computer and robotic agency and their capabilities.

4.4 The Collaborative Team Model

4.4.1 Workload Results

The scales for each IMPRINT Pro workload channel differ. The Auditory, Cognitive and Fine Motor channels are scaled from one to eight. The Visual and Gross Motor channels are scaled from one to seven and the Tactile and Speech channels are scaled from one to five. The model outputs the workload values for each channel at each time point. Fine motor and gross motor workload have been combined in order to permit a comparison with participants' in situ subjective workload ratings. The total modeled workload is calculated by rescaling each channel to a one to five scale, summing the channels at each time point and calculating a time-weighted mean for each of the six Investigation Areas. The minimum possible workload is six and the maximum workload value is 30. Each model was run ten times and the means for each area workload value were taken over those ten runs.

The mean total workload modeled per area for the H-H model was 13.18 (St. Dev. = 0.58) and was 12.96 (St. Dev. = 0.47) in the H-R model. A Kruskal Wallis test indicated that total workload in the H-H model was significantly higher than total workload in the H-R model, $\chi^2(1) = 9.02$, $p = 0.003$.

The mean total modeled workload across both models by Investigation Index are presented in Table 7. A Kruskal Wallis test indicated a significant main effect of Investigation Index on total modeled workload, $\chi^2(2) = 7.015$, $p = 0.030$. Mann Whitney U tests, with a Bonferroni adjusted $\alpha = 0.0167$, revealed no significant comparisons between Investigation Index.

The mean total modeled workload by condition and Investigation Index are presented in Table 7. A Kruskal Wallis test indicated a significant interaction effect of condition and Investigation Index on total modeled workload, $\chi^2(5) = 16.998$, $p = 0.005$ independent of condition. Mann Whitney U tests, with a Bonferroni adjusted $\alpha = 0.0033$, revealed no significant comparisons within conditions or between conditions for the same Investigation Index.

Table 7. Mean Modeled Workload by Investigation Index and condition. Standard Deviations are listed in parentheses.

Investigation Index	Across Both Models	H-H Model	H-R Model
Low	12.78 (0.71)	12.86 (0.80)	12.70 (0.61)
Medium	13.24 (0.33)	13.41 (0.31)	13.08 (0.27)
High	13.19 (0.36)	13.28 (0.35)	13.10 (0.35)

The mean total modeled workload by condition is presented in Table 8. A Kruskal Wallis test indicated a significant main effect of Investigation Area on total modeled workload, $\chi^2(5) = 96.13$, $p < 0.001$. Pairwise Wilcoxon rank sum tests with Holms p-value correction indicated a set of significant differences. Area 1 yielded significantly lower workload than all other areas ($p < 0.001$ in all cases). Investigation Areas 2, 3 and 4 led to higher workload than Areas 5 and 6 ($p < 0.001$ in all cases).

The mean total modeled workload by condition and Investigation Area is presented in Table 8. A Kruskal Wallis test indicated the presence of a significant interaction effect of condition and Investigation Area on total modeled workload, $\chi^2(11) = 108.93$, $p = 0.001$, independent of condition. Pairwise Wilcoxon rank sum tests with Holms p-value correction indicated a set of significant differences.

Comparisons between areas within only the H-H model indicated Area 1 (Low Investigation Index) had significantly lower workload than all other areas ($p = 0.011$ in all cases). Area 5 (Medium Investigation Index) had significantly lower workload than Areas 2 (Medium Investigation Index) and 4 (High Investigation Index) ($p = 0.011$ in both cases). Area 6 (High Investigation Index) had significantly lower workload than Areas 2, 3 (Low Investigation Index), 4 and 5 ($p = 0.011$ in all cases). Comparisons within the H-R model indicated that Area 1 had significantly lower workload than all other areas ($p = 0.011$ in all cases). Areas 5 and 6 both had significantly lower workload than Areas 2, 3 and 4 ($p = 0.011$ in all cases). Comparisons of workload between the two models demonstrated that the H-H model had significantly higher workload in Area 2 ($p = 0.043$), Area 4 ($p = 0.043$), Area 5 ($p = 0.011$) and Area 6 ($p = 0.011$).

Table 8. Mean Modeled Workload by Investigation Area both Across both models and by condition. Each area's associated Investigation Index is provided. Standard Deviations are listed in parentheses.

Investigation Area	Across Both Models	H-H Model	H-R Model
Area 1 - Low	12.10 (0.09)	12.10 (0.09)	12.10 (0.09)
Area 2 - Medium	13.48 (0.27)	13.65 (0.24)	13.31 (0.17)
Area 3 - Low	13.46 (0.25)	13.62 (0.26)	13.29 (0.05)
Area 4 - High	13.51 (0.15)	13.60 (0.17)	13.42 (0.03)
Area 5 - Medium	13.00 (0.17)	13.16 (0.06)	12.84 (0.05)
Area 6 - High	12.87 (0.15)	12.97 (0.04)	12.78 (0.17)

Overall, workload was significantly higher in the H-H model than in the H-R model. While there were significant effects of Investigation Index and an interaction effect between condition and Investigation Index, individual comparisons did not reveal any significant results. Analysis by area showed that Areas 1, 5 and 6 tended to have lower workload than Areas 2, 3 and 4. The H-H model had

significantly higher workload in all areas except for Areas 1 and 3, the two areas with a Low Investigation Index. Consideration of the empirical data should include the important trends present in the model: a lower level of workload in the H-R model and no manipulation of workload by Investigation Index.

4.4.2 Reaction Time Results

The Collaborative Team model incorporated a representation of Secondary Task Question response time and Item Reaction Time. The time it took to respond to an On List question in the H-H model was based on the time it takes to comprehend a question after the asker finishes speaking (0.72 s), decide if a name matches the given list (0.45 s), and say a one-word response (0.34). The response times are estimated to be 1.51 s. The H-R model took 0.21 s longer for the participant to finish processing the robot's question for a predicted On List response time was 1.72 s.

Danger Level questions in the H-H model are estimated to incorporate the time it takes to comprehend a question after the asker finishes speaking (0.72 s), matching the name to the associated Danger Level (0.38) and speaking a one-word response (0.34). The estimated Danger Level response time is 1.44s. Again, the H-R model predicted that participants require 0.21 s longer to process the robot's question, so the predicted Danger Level response time was 1.65.s.

Based on IMPRINT Pro's micromodels of behavior [41], the time required to react to a potentially out of place item in the environment was 1.41 seconds. This time was based on the combination of eye fixation time (0.30 s), eye movement time (0.10 s), head movement time (0.20 s), decision time (0.07 s) search time in the environment (0.60 s) and the time to determine if an item was out of place (0.24 s).

4.4.3 Discussion

The models provided information regarding predictions of workload and reaction time in both the H-H and the H-R conditions. Any differences in workload between the two models can be explained by the lengthening of robot speech and movement tasks in the model. Empirical results reflecting workload levels and timing information corroborate that the difference in workload between the two conditions is based upon the longer time a robot took to move and speak (see Section 4.5).

Given model results, it is apparent that Low, Medium and High Investigation Indices do not necessarily correspond to Low, Medium and High levels of workload. Investigation Indices were calculated by enumerating the amount of work a participant was to complete in each of the Investigation Areas. What was not completely accounted for in these calculations was that a larger amount of work takes a longer amount of time to complete, potentially with the same proportion of work to time in all three Investigation Indices. Therefore, this issue creates a potentially inconsistent way to manipulate workload. Yet, analysis of the evaluation results are analyzed by Investigation Index to determine which workload measures may reflect total work or a longer time, rather than actual workload levels. The Low Investigation Index areas had lower workload (though not significantly), but the difference between Medium and High Investigation Indices is unclear.

The item reaction time results provide an estimation of how long it will take participants to respond to a potentially out-of-place item in their field of view. Reaction time measurements are typically made in controlled environments with a limited set of options requiring a response. This model's aim was to compare modeled reaction time to the validation evaluation results.

The validation experiment results confirm a) the workload in the H-R condition is lower than the H-H team, b) the lower workload is due only to slower robot timing, c) that there is an effect of Investigation Index, even if it may be an inconsistent measure of workload and d) validation of item reaction time.

4.5 Collaborative Team Validation Evaluation

The validation of the model results was executed by comparing trends from the validation evaluation results to those predicted by the model. Results from the model validation experiment, in general, are analyzed by condition, Investigation Index and the interaction between the two, with the exception of results that do not have measurements during each Investigation Area (items assessed, total investigation time, NASA-TLX), which are analyzed by condition.

4.5.1 Workload Results

4.5.1.1 Physiological Measures

The mean HRV for the H-H condition participants was 1.60 (St. Dev. = 1.82) and 1.46 (St. Dev. = 1.43) for the H-R condition participants. The mean HRV in the Low Investigation Index areas was 1.59 (St. Dev. = 1.59), 1.55 (St. Dev. = 1.88) in the Medium Investigation Index areas and 1.45 (St. Dev. = 1.40) in the High Investigation Index areas. No significant main effect of condition, main effect of Investigation Index or interaction effect of condition and Investigation Index on HRV was found. This result does not correspond with the model's prediction of lower workload and, therefore lower HRV, in the H-R condition.

The mean normalized heart rate for the H-H condition was -3.15 (St. Dev. = 26.95) beats per minute and 3.83 (St. Dev. = 10.69) beats per minute for the H-R condition. H-H participants did tend to have negative values for normalized heart rates, while H-R participants tended to have positive values. This result shows that the H-H condition participant heart rates tended to be lower than their base rate and H-R condition participants tended to have more elevated heart rates from their base heart rates. The mean normalized heart rate in areas with a Low Investigation Index was 1.63 (St. Dev. = 20.93) beats per minute, 0.03 (St. Dev. = 20.36) beats per minute in Medium Investigation Index areas and -0.64 (St. Dev. = 21.17) beats per minute in High Investigation Index areas. No significant main effect of condition, main effect of Investigation Index or interaction effect of condition and Investigation Index on normalized heart rate was found. This result also does not correspond with the model's prediction of lower workload and, therefore, lower heart rate in the H-R condition.

The mean normalized respiration rate for H-H condition participants was -0.07 (St. Dev. = 3.62) breaths per minute and 2.54 (St. Dev. = 4.03) breaths per minute for the H-R condition. A Kruskal

Wallis test indicated that the mean normalized respiration rate was significantly higher during the H-R condition, $\chi2(1) = 21.45$, $p < 0.01$. The mean normalized respiration rate by condition and Investigation Index is presented in Table 9. A Kruskal Wallis test indicated that there was no significant main effect of Investigation Index. A Kruskal Wallis test indicated that there was a significant interaction effect of condition and Investigation Index on normalized respiration rate, $\chi2(5) = 22.13$, $p < 0.01$. A series of Mann Whitney U tests with a Bonferroni adjustment of $\alpha = 0.0$ found that H-H participants had a significantly lower normalized respiration rate for the High Investigation Index area than their counterparts in the H-R participants, $U = 380$, $Z = -3.02$, $p = 0.002$.

Overall, participants in the H-R condition had higher Normalized Respiration Rate than participants in the H-H condition. This difference was most pronounced in areas with a High Investigation Index. This data corresponds with literature that indicates a negative correlation between workload and respiration rate, meaning that when respiration rate increases, workload decreases [61].

Table 9. Mean Normalized Respiration Rate by condition and Investigation Index.

Investigation Index	Across Both Conditions	H-H	H-R
Low	1.40 (4.08)	0.20 (3.62)	2.60 (4.18)
Medium	0.99 (4.13)	-0.27 (3.77)	2.25 (4.14)
High	1.31 (3.96)	-0.14 (3.56)	2.76 (3.86)

4.5.1.2 Investigation Timing

The mean total time spent investigating the entire area during the H-H condition was 1964.18 (St. Dev. = 373.79) seconds (s) and 2473.90 (St. Dev. = 213.16) s in the H-R condition. A Kruskal Wallis test revealed that H-R condition participants took significantly longer, which may be attributed to the longer time it took the robot to speak and travel down the hallway, $\chi^2(1) = 14.58$, $p < 0.01$. The fact that the H-R condition required significantly longer investigation time is not surprising, as the robot took longer to do some of the tasks than the human partner. However, lower levels of workload in the H-R condition can be attributed to the participants performing the same amount of work as H-H condition participants, just in a longer amount of time. Answering this question involves testing the strength of the correlation between these longer investigation times and the in situ subjective workload ratings (See Sections 4.5.1.5 and 4.5.1.8).

4.5.1.3 Secondary Task Questions

A total of 172 correct and 37 incorrect responses to the secondary task questions were provided during the H-H condition. The H-R condition participants provided 156 correct and 54 incorrect responses. The number of responses does not match across conditions, because the experimenter neglected to ask six questions total in the H-H condition and four questions total in the H-R condition. A Pearson's Chi-squared test with Yates' continuity correction found no significant main effect of condition for the On List Correct metric.

The number of correct and incorrect On List responses by condition and Investigation Index are provided in Table 10. The On List question responses, independent of condition were analyzed by

Investigation Index (Section 4.3.1) in order to determine whether secondary task question responses were impacted by the environmental based workload manipulation. A Chi-squared test indicated a significant main effect of Investigation Index on the number of correct responses, $\chi^2(2) = 9.57$, $p < 0.01$. Three Mann-Whitney comparisons were performed, with a Bonferroni adjustment for family-wise error in order to determine which Investigation Indices were different, making the requirement for significance $\alpha = 0.017$. The number of correct responses for the Low Investigation Index was significantly higher than that for the High Investigation Index, $U = 11573$, $Z = -2.930$, $p < 0.01$. No other comparisons were significant.

A Pearson's Chi-squared test with Yates' continuity correction found a significant interaction effect of condition and Investigation Index for the correct On List responses, $\chi^2(11) = 152.28$, $p < 0.01$. Mann-Whitney pairwise comparisons were performed with Bonferroni family-wise adjustments, $\alpha = 0.0033$. The Low Investigation Index responses in the H-R condition were correct significantly more frequently than the High Investigation Index areas in the H-R condition, $U = 2730$, $Z = -3.94$, $p < 0.001$.

Table 10. Total number of correct and incorrect responses to On List questions, by Investigation Index and condition.

Investigation Index	Across Both Conditions		H-H		H-R	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Low	119	21	55	15	64	6
Medium	111	28	61	8	50	20
High	98	42	56	14	42	28

The On List Correct results demonstrate that workload reserves were not significantly different between conditions. Additionally, the number of correct responses was significantly lower in areas with High Investigation Areas than the Low Investigation Index areas. While this result offers evidence that workload reserve levels were manipulated by Investigation Index, the data do not fully support this claim. The H-R condition participants were dramatically affected by the changes in Investigation Index. The H-H condition participants' correct responses were not significantly affected by the investigation index.

The secondary task questions incorporate a second question requiring the participants to provide the danger level associated with a chemical when the participants indicated that a chemical was on the provided list (whether the response was correct or not). The number of correct and incorrect responses to the Danger Level questions, by condition and Investigation Index are provided in Table 11. The H-H condition participants provided 62 correct and 39 incorrect responses, while the H-R condition participants had 65 correct and 67 incorrect responses. During the H-H condition, the danger Level questions were asked 101 times and 132 times in the H-R condition. The effect of the Pearson's Chi-squared test with Yates' continuity correction found no significant main effect of condition on the Danger Level Correct metric.

A significant main effect of Investigation Index, independent of condition was found by the Pearson's Chi-squared test with Yates' continuity correction, $\chi^2(2) = 11.13$, $p < 0.01$. A Bonferroni-adjusted Mann Whitney U test comparison indicated that there were a significantly higher number of correct responses in the Low Investigation Index areas than in areas with a High Investigation Index, U = 2265, z= -3.23, p = 0.002. No other pairwise comparisons were significant.

A Pearson's Chi-squared test with Yates' continuity correction found a significant interaction between condition and Investigation Index, $\chi^2(11) = 41.87$, $p < 0.01$. Mann Whitney U test pairwise comparisons with Bonferroni correction for family-wise error indicated that participants in the H-R condition responded with a higher rate of correctness in areas with a Low Investigation Index than in areas with a High Investigation Index, U = 1145, z = -3.40, p < 0.001. No other pairwise comparisons were significant.

Table 11. Total number of correct and incorrect responses to Danger Level questions, by Investigation Index and condition.

Investigation Index	Across Both Conditions		H-H		H-R	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Low	67	35	31	16	36	19
Medium	41	40	20	14	21	26
High	19	31	11	9	8	22

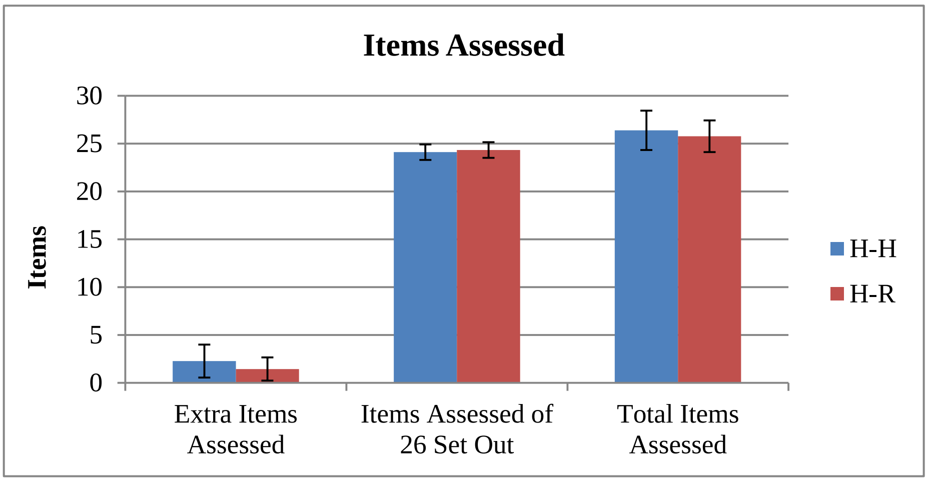
Overall, there was not a significant main effect of condition on the number of correct responses to Danger Level questions. The participants in both conditions correctly answered the Danger Level questions with a similar frequency. There was a significant main effect of Investigation Index, indicating that overall participants responded with correct answers more often in areas with a Low Investigation Index than in areas with a High Investigation Index, which was an evident trend for the H-R condition participants. There were no significant differences between Medium Investigation Index responses and either of the other two Investigation Indices. This result reflects only partial support of a manipulation of workload reserves by Investigation Index. Differences between all three of the Investigation Indices were not significant, but the general trend in the data shows the number of correct responses decreased as the Investigation Index increased from Low to Medium to High.

The secondary task question results overall indicate that participants from both conditions did not have a significantly different number of correct responses. The main effect of Investigation Index showed in both the On List question responses and the Danger Level responses that there was a higher frequency of correct responses in Low areas than in High areas. This result reflects the trend of lower modeled workload in Low Investigation Index areas than in High Investigation Index areas.

4.5.1.4 Items Assessed

The number of items participants found by condition is presented in Figure 17. The number of items found are presented by the total out of the 26 items placed by experimenters (please see detailed list in Section 4.3.4), the extra items assessed, and the total set of items found.

Figure 17. The mean number of Items assessed by participants in each condition. Error bars represent



one standard deviation above and below the mean

The H-H condition participants assessed an average of 24.11 (St. Dev, = 0.81) items, while the H-R condition participants assessed an average of 24.33 items (St. Dev. = 0.82). A one-way ANOVA found no significant effect of condition for the number items found.

Some participants pointed out additional items that they suspected may be related to the bomb threat. These extra items included signs, bulletin board postings and items such as paint cans or rubber boots that were not placed in the environment by the experimenters. The H-H condition participants found a mean of 2.28 extra items (St. Dev. = 1.73), while the H-R condition participants found a mean of 1.44 extra items (St. Dev. = 1.21). A one-way ANOVA found no significant effect of condition for extra items found. Overall, there was not a significant difference in the number of items assessed by participants in either condition.

4.5.1.5 Subjective Workload Ratings

Total subjective workload was the sum of ratings from 1 to 5 on the six workload channels, thus the possible total workload ratings ranged from 6 to 30. The median total workload rated by the H-H participants was 13 and the median for the H-R participants was 11. A Kruskal-Wallis test showed that the H-H participant ratings were significantly higher than the H-R condition ratings, $\chi2(1) = 18.64$, $p < 0.01$.

The median total workload rated while assessing areas with a Low Investigation Index was 11.5. While assessing areas with a Medium Investigation Index, participants rated total workload with a median of 10.5 and in the High Investigation Index areas participants rated workload with a median of 12.5. A Kruskal-Wallis test indicated no significant effect of Investigation Index on total subjective workload ratings.

The median total workload ratings for each Investigation Index in both conditions are provided in Table 12. A Kruskal-Wallis test indicated a significant interaction effect of Investigation Index and condition on total Subjective Workload Ratings, $\chi2(5) = 22.22$, $p < 0.01$, independent of condition. A series of Mann Whitney U tests were performed and with Bonferroni adjustments, $\alpha = 0.0033$ to

determine which interactions were significant. None of the median ratings were significantly different between Investigation Indices within both the H-H and H-R conditions.

Overall, the H-H condition participants rated their workload higher than the H-R participants. There was no significant main effect of Investigation Index. The H-H condition participants rated workload significantly higher than H-R condition participants in the areas with a Medium Investigation Index. There were no significant differences between ratings from each Investigation Index within each condition.

Table 12. Median total Subjective Workload Ratings by condition and Investigation Index.

Investigation Index	H-H	H-R
Low	12.0	11.0
Medium	13.0	9.0
High	14.5	12.0

4.5.1.6 Comparing Modeled Workload and In Situ Workload Ratings

The total modeled workload is calculated based upon the same total workload scale as the in situ workload ratings, thus the scores can be directly compared. The mean total subjective workload rating in the H-H condition was 13.99 (St. Dev. = 5.24) and 10.97 (St. Dev. = 3.98) in the H-R condition. The mean total modeled workload was 13.07 (St. Dev. = 0.53), while the mean total subjective workload rating was 12.48 (St. Dev. = 4.88). A Kruskal Wallis test indicated that the modeled workload was significantly higher than subjectively rated workload, $\chi^2(1) = 15.08$, $p < 0.001$.

Workload was compared between the model and the subjective ratings by Group: the H-H condition, the H-R condition, the H-H model and the H-R model. A Kruskal Wallis test indicated a significant main effect of Group, $\chi^2(3) = 38.89$, $p < 0.001$. Pairwise Wilcoxon rank sum tests with Holms p-value correction indicated a set of significant differences. The H-H condition subjective ratings were significantly higher than the H-R condition subjective ratings ($p < 0.001$), as seen when the model was analyzed on its own in Section 4.4.1. The H-H model workload was significantly higher than the H-R model workload ($p = 0.008$). The H-H condition subjective ratings and the H-H model workload were not significantly different. The H-R condition subjective ratings were significantly lower than the H-R model workload ($p < 0.001$).

Workload by investigation area was compared between modeled results and subjective workload ratings. Figure 18 presents the workload by each group (H-H condition subjective ratings, H-H model workload, H-R condition subjective ratings and H-R model workload) and area. A Kruskal Wallis test indicated a significant interaction effect between Area and Group on workload, $\chi^2(23) = 68.88$, $p < 0.001$. Pairwise Wilcoxon rank sum tests with Holms p-value correction indicated a set of significant differences, but no significant difference was found between the H-H condition subjective ratings and the H-H model results or between the H-R condition subjective ratings and the H-R model results.

Modeled workload results were significantly higher than subjective workload ratings, particularly when comparing the H-R condition subjective ratings and the H-R model workload. However, when comparing workload between groups by area, there were no significant differences.

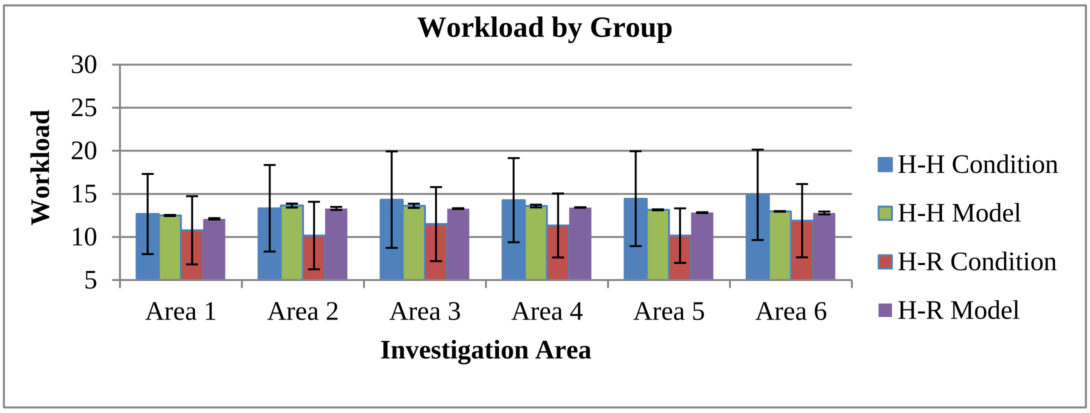


Figure 18. Workload by Area and Group.

4.5.1.7 NASA-TLX Responses

The mean total NASA-TLX score for the H-H participants was 41.32 (St. Dev. = 17.80) and 30.89 (St. Dev. = 15.12) for the H-R participants. A two-sided t-test indicated no significant difference, but the H-H responses tended to be higher. Figure 19 provides the mean responses to each NASA-TLX workload component. A series of t-tests compared the responses from each of the workload component between conditions and resulted in no significant differences.

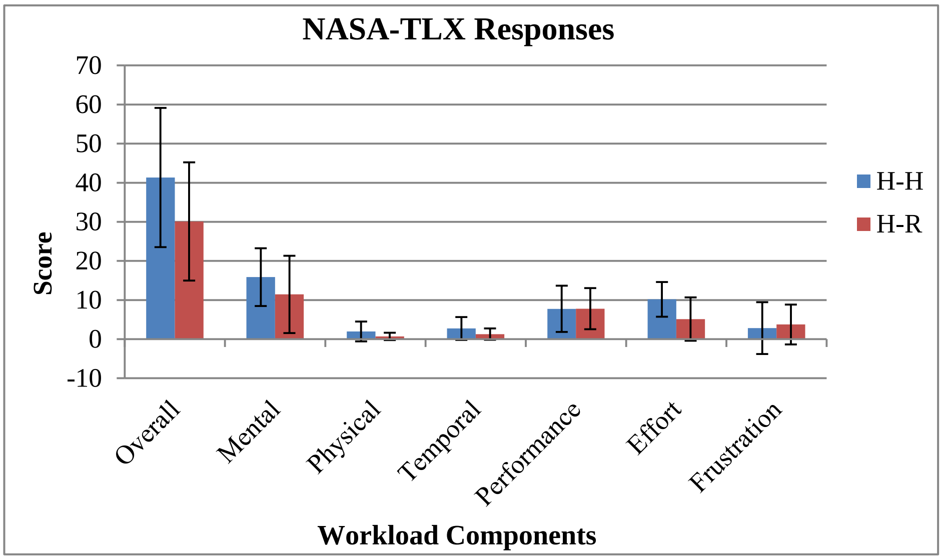


Figure 19. Responses along workload components for NASA-TLX Scores.

4.5.1.8 Correlations Analysis

It is important to understand the relationships between data sets, thus Pearson's product-moment correlations were performed to correlate physiological data with the Total Subjective Workload Ratings. The model predicted that H-R condition workload is lower simply due to timing of robot speech and movement. The in situ workload ratings confirmed that the H-R condition workload was

indeed lower, but the comparison between the model and the evaluation results indicated that the gap between modeled workload and rated workload was larger in the H-R condition. Determining whether higher workload ratings in the H-H condition were simply due to the shorter time frame in which H-H condition participants completed tasks necessitated that area investigation time was correlated to total subjective workload ratings. Timing and Total Subjective Workload Ratings were not significantly correlated to one another.

HRV and Normalized Heart Rate were not significantly affected by either condition or Investigation Index and these physiological data potential correlations to total subjective workload ratings were not tested. Normalized Respiration Rate was significantly higher in the H-R condition, notably in areas with High Investigation Indices (See Section 4.5.1.1). There was a significant negative correlation between Normalized Respiration Rate and Total Subjective Workload Ratings, $r(214) = -0.17$, $p = 0.01$. This result indicates that when Total Subjective Workload Ratings were high, Normalized Respiration Rate was low.

4.5.2 Discussion of Workload Validation Evaluation Results

The hypothesis that the models are a good predictor of workload in both experimental conditions was supported for the H-H condition. There was no significant difference between the H-H model and the H-H condition in situ workload ratings. The workload values produced by the H-R model; however, were significantly higher than the H-R condition results even though the model accounted for the robot's slower speech and movement.

The hypothesis was that workload is lower for the H-R condition was supported. The model data and the in situ workload ratings were both found to be lower and the respiration rate data had a significant negative correlation to in situ workload ratings.

One method of computing workload calculates the proportion of work completed to the time taken to complete it. Participants did not assess a significantly different number of items in either condition, representing the amount of work completed. Participants in the H-R condition took significantly longer to complete the investigation. Calculating this proportion of work to time taken to complete the evaluation results in lower workload in the H-R condition. This result was found in the model data. A confirmation of the lower workload resulting from the longer time taken to complete the tasks in the evaluation requires a significant correlation between investigation time and in situ workload ratings, which did not exist. Therefore, the longer time taken for H-R condition participants to complete the evaluation was not the only contributing factor to their lower workload.

The nature of the collaborative relationship between the human and robot in the H-R condition seemingly lowered workload. One theory is that participants may have overestimated the robot's capabilities. Participants may have felt less pressure, assuming that the robot would correct any mistake they made. The participants were told that the robot was completely autonomous, thus the lack of human involvement in judging their performance may have alleviated pressure and led to the participants rating their workload lower.

Another theory is that the participants perceived working with the robot was perceived as fun. The participants' low level of experience working with robots prior to the evaluation and the experience of working with a robotic teammate may have been perceived as fun or interesting. The participants paired with a human teammate may have been more focused on the nature of the scenario – a bomb threat – and were able to perceive the work being done as real work. Evidence for this hypothesis is present in the evaluation videos; participants in the H-R condition appeared to have more instances of lighthearted tone. Future work includes enumerating these comments for each condition will analyze this hypothesis.

This research makes contributions to the human-robot interaction field by demonstrating that workload is lower for H-R teams and that this result may not be attributed to H-R teams requiring more time to complete the tasks. The question of why this discrepancy occurs reflects the nature of the human-robot collaborative relationship. Future avenues of applying this work include investigating how degrees of assumptions of robot capability affect human-robot collaboration and workload. The lower perception of workload by H-R condition teams may indicate a benefit of working in a collaborative relationship with a robot. However, it may also reflect a lower level of taking the task seriously, which can impact mission success. This research poses the question of whether or not lower levels of workload exist for different human-robot team relationships and why.

4.5.3 Comparison Workload across Teaming Evaluations

A comparison of the results from the Guided Interaction Team model and evaluation to the results of the Collaborative Team model and evaluation are presented. It is important to compare the results from the two evaluations in order to determine whether the nature of the collaborative relationship between the human and the robot evaluation alters the conclusions drawn from the Guided Interaction Team evaluation.

4.5.3.1 Guided Interaction Team Total Modeled Workload

This analysis presents total modeled workload values in the Guided Interaction Team scenario that were calculated from the original scales in IMPRINT Pro to match the same scales as the in situ workload ratings. Previous analysis of Guided Interaction Team model means employed re-scaled the in situ workload data to match the IMPRINT Pro workload scales, but in order to compare these results to the Collaborative Team evaluation results, all modeled workload is converted to a scale of one to five for each channel. The Guided Interaction Team model had no built-in uncertainty and the modeled workload has no standard deviation by triage assessment. However, the presented means have standard deviations associated with the means of multiple victim assessment workload values. The mean workload in the H-H model across the 11 triage assessments was 15.73 (St. Dev. = 2.36) and was 15.17 (St. Dev. = 1.74) in the H-R model. A Kruskal Wallis test indicated that there was no significant difference between the models. The mean modeled workload for Delayed victims was 14.80 (St. Dev. = 2.39), 16.07 (St. Dev. = 2.04) for Immediate victims and 14.58 (St. Dev. = 1.07) for Expectant victims. A Kruskal Wallis test indicated no significant main effect of triage level. The

model predictions indicated that the Guided Interaction Team H-H condition workload may be slightly higher than the H-R condition workload, but the lack of statistical significance prevents strong assertions.

4.5.3.2 Physiological Measures

The Guided Interaction Team validation evaluation analysis [37] determined that heart rate and respiration rate were significantly higher in the H-H condition, but were not significantly affected by triage level. HRV showed no differences between conditions, but a main effect of triage level (See Section 3.2).

The Collaborative Team evaluation results found that heart rate was not significantly different between conditions and respiration rate was higher in the H-R condition (Section 4.5.1.1). These results are inconsistent with the Guided Interaction Team findings. Both evaluations found no significant difference between conditions for HRV. The results from the Guided Interaction Team evaluation indicate that HRV was affected by a main effect of Triage Level, but a similar finding was not supported by either heart rate or respiration rate data in the Guided Interaction Team evaluation. The lack of consistent effects of triage level shows that workload manipulations by Triage Level were not measurable using physiological data in the Guided Interaction Team evaluation.

4.5.3.3 Timing

Timing was calculated by measuring the amount of time it took for participants to perform each triage assessment. The Guided Interaction Team evaluation H-R condition participants took significantly longer to perform triage assessments and the Guided Interaction Team timing results are very similar to those from the Collaborative Team evaluation. Namely, the H-R participants take longer to complete tasks and timing is dependent on the amount of work to be done for an individual triage assessment or investigation area.

4.5.3.4 Secondary Task Questions

The analysis of the Guided Interaction Team evaluation Secondary Task Question responses (presented in Section 3.4.3 indicated that there were no significant differences between the two conditions for Secondary Task Question correctness. Both scenarios show that participants in either condition do not respond significantly differently to the secondary task questions. The correctness of responses did not depend on the partner condition.

4.5.3.5 Task Performance

The Guided Interaction Team evaluation required participants to report each victim's age and the breathing rate during 9 of the 11 triage assessments. During the H-H condition, the reported ages of Victim 1 were not recorded, so reported victim ages were only analyzed between the H-H and H-R conditions for Victims 2 through 6. The mean reported ages of these victims by condition are presented in Table 13. T-tests compared the responses between conditions and found no significant differences.

Table 13 Mean Reported Age by condition and Victim number, in years.

Victim Number	H-H	H-R
2	33.21 (3.72)	33.64 (6.21)
3	6.32 (1.73)	7.50 (1.87)
4	23.89 (6.57)	22.93 (4.45)
5	0.92 (0.53)	1.32 (0.97)
6	31.46 (4.64)	33.29 (6.34)

The reported breathing rates of victims were also compared for each victim assessment using t-tests between conditions. The mean reported breathing rates by condition are presented in Table 14. Victim Assessment 8 was the only assessment demonstratig a significant difference between the two conditions. Participants in the H-H condition reported the victim's breathing as significantly higher than the H-R participants, $t(25) = 2.50$, $p = 0.02$. The delta between the two mean reported respiration rates was 2.42 breaths per minute.

Table 14 Mean Reported Respiration Rate by condition and Assessment number, in breaths per minute.

Assessment Number (of 11)	H-H	H-R
4	20.79 (0.97)	20.80 (1.03)
5	21.08 (2.23)	19.44 (0.73)
6	27.00 (2.91)	26.00 (1.29)
8	20.57 (1.45)	18.15 (3.29)
9	24.93 (6.68)	21.38 (7.95)
10	27.57 (1.50)	28.86 (9.69)
11	14.38 (8.68)	11.89 (1.69)

These results represent an aspect of performance in the Guided Interaction Team evaluation. The number of items found in the Collaborative Team evaluation represents task performance. A lower number of items found reflects that participants were not performing the task of assessing items well and were missing some of the items. Performance in the Guided Interaction Team evaluation cannot directly crossover to number of victims assessed, because participants were following their partner's instructions to perform each of the eleven victim assessments. Performance can be determined by assessing the quality of the victim assessments, reflected by the information given about each of the victims by the participant. If the results are not significantly different, a similar level of performance on the task to report information can be assumed. Comparisons between conditions for reported breathing rates and victim ages were not significantly different (with one exception). This result indicates that the H-H and H-R participants in the Guided Interaction Team evaluation had similar levels of performance.

The Collaborative Team evaluation's performance metric included the number of items assessed and there was no significant difference between the items assessed for the two conditions Therefore, results from both evaluations support the point that participants did not perform significantly differently on the mission by condition.

4.5.3.6 Subjective Workload Ratings

The Guided Interaction Team evaluation total subjective workload rating results (See Section 3.4.2) demonstrate that workload was significantly lower in the H-R condition. A main effect of Triage

Level demonstrated the workload manipulation effect. Workload ratings from the Collaborative Team evaluation also showed that H-R condition workload was lower, but did not demonstrate a full manipulation of workload by Investigation Index.

4.5.3.7 NASA-TLX Responses

There was not a significant difference between the two conditions in either evaluation, but both sets of results showed lower total workload scores in the H-R condition [37]. Guided Interaction Team evaluation results are presented in Section 3.4.4.

4.5.3.8 Comparing Modeled Workload and In situ Workload Ratings

Modeled workload values presented in Section 3.2 were based on IMPRINT Pro's workload scales. The same model results were converted to the in situ workload rating scales (one to five for each channel) for comparison to the collaborative team model. The modeled workload values, after conversion, range from 6 to 30 for direct comparison with the in situ model ratings. After conversion, the mean of the modeled workload across both conditions in the guided interaction team was 15.45 (St. Dev. = 2.04) and was 14.87 (St. Dev. = 5.77) across both conditions for the in situ workload ratings. New analysis using a Kruskal Wallis test indicated that there was no significant difference between the modeled workload across both conditions and the in situ workload ratings.

This analysis uses the word Group to include the H-H model, the H-R model, the H-H condition and the H-R condition workload values. A Kruskal Wallis test indicated there was a significant effect of Group on workload values, $\chi^2(3) = 15.15$, $p = 0.002$. Post hoc pairwise Wilcoxon tests indicated that the only significant difference was that the H-H participants rated workload significantly higher than the H-R participants ($p = 0.002$). There were no significant differences between the H-H model and the H-H condition, between the H-R model and the H-R condition, or between the H-H model and the H-R model. Both Guided Interaction Team models appeared to be good predictions of workload results, despite the fact that the evaluation workload values were not significantly different between the models. This result differs from the Collaborative Team evaluation, because the Collaborative Team models predicted significantly higher workload than was found from the in situ workload ratings.

4.5.3.9 Correlations Analysis

Results from the Guided Interaction Team evaluation yielded a significant positive correlation between heart rate and total in situ workload ratings, and a significant negative correlation between respiration rate and total subjective workload ratings (See Section 3.4.5).

While the correlation between normalized heart rate and total subjective workload ratings was not relevant in the Collaborative Team evaluation, a significant negative correlation between normalized respiration rate and total in situ workload ratings was also found.

An additional Pearson's product-moment correlation was performed on Guided Interaction Team evaluation data and found a significant correlation between time taken to assess each victim and in

situ workload ratings, r(293) = 0.16, p = 0.005. This result indicates that the lower in situ workload ratings may be due to the fact that the triage assessments took a longer time in the H-R condition. This result is opposite from the same correlation for the Collaborative Team evaluation.

4.5.3.10 Discussion of Comparison across Teaming Scenarios

The Guided Interaction Team model workload values were not significantly different from the in situ workload ratings. The Guided Interaction Team evaluation indicated that while participants rated workload significantly lower in the H-R condition, a significant correlation between timing data and in situ workload ratings may explain the difference. The models provided a good prediction of workload, despite not indicating the significant difference between the two conditions.

Data from both evaluations support the participants' ability to attain a similar performance level for assigned tasks, regardless of condition. This result is supported in the Guided Interaction Team evaluation by the secondary task question responses, the reported victim ages and the reported victim respiration rates. The Collaborative Team evaluation results also supported that performance levels were similar between the two conditions.

The lower total subjective workload ratings and NASA-TLX scores in the H-R condition of the Collaborative Team evaluation shadow the subjective results from the Guided Interaction Team. Participants perceived less workload in the H-R condition despite a lack of training in robotics or first response. Having a human partner does not seem to offer enough reassurance in an uncertain situation to lower workload. Physiological results remain unclear on their own, but the negative correlation between Total Subjective Workload Ratings and normalized respiration rate can be useful when analyzing live data – a sudden drop in respiration rate can indicate an increase in workload.

Overall when comparing data across the two evaluations, it is apparent that a) workload ratings are lower in the H-R condition, b) task performance is not different between the two conditions and c) H-R teams take longer to complete the given task. A discrepancy between the two evaluations is why workload is lower in the H-R condition. The Guided Interaction Team evaluation's data indicates that the possible reason for the lower H-R condition workload is the longer time participants took to perform the tasks. This conclusion cannot be drawn in the Collaborative Team evaluation. The differences in workload in the Collaborative Team evaluation can be attributed to the presence of a collaborative relationship between the human and robot which did not exist in the Guided Interaction Team, as discussed in Section 4.5.2.

4.5.4 Workload Discussion

Prior research investigated workload in a Guided Interaction Team with human-robot teams that did not incorporate any collaboration and an instruction-based relationship between teammates [37, 38]. The aim of the comparison between data from the instruction-based Guided Interaction Team evaluation and the Collaborative Team evaluation was to determine whether the conclusions drawn from the Guided Interaction Team evaluation hold with the addition of a collaborative relationship in

the Collaborative Team evaluation. The relationship in the Collaborative Team evaluation was achieved by assigning co-located partners to work together on the same task. Multiple measurements of workload were recorded.

The relationship between the human and robot in the Collaborative Team evaluation successfully fulfilled Bratman's three characteristics of a collaboration, which include responsiveness of teammates, commitment to the activity and commitment to support of the team [13]. The mutual responsiveness to a partner was achieved by creating a dynamic system that addressed participant questions and offered feedback. The commitment to the joint activity was achieved by having each partner responsible for a distinct set of tasks; for example, the participant was responsible for reading signs while the robot was responsible for investigating each fire extinguisher. The mutual support of a teammate was demonstrated by having the teammates discuss each step with each other and make joint decisions together regarding next steps in the investigation. The collaboration between the two team members allowed for a successful investigation. This relationship was more flexible than the relationship fostered in the Guided Interaction Team.

However, the main finding in the Collaborative Team evaluation results shows that timing is not the only reason that workload was lower in the H-R condition. The main difference between the two scenarios is nature of the collaborative relationship between a human and robot. This difference in interaction appears to have lowered workload in the H-R condition, which was not included in the model. Future work can involve determining the proportion of the difference in workload attributed to timing and human-robot collaboration.

The hypotheses stated that the model will be a good predictor of workload in both evaluation conditions and workload will be lower in the human-robot team. Workload was definitely lower in the human-robot teams, but the H-R model was unable to create an accurate prediction of the workload levels. In the Guided Interaction Team, the model's workload predictions were not significantly different from the in situ workload ratings. However, rather than simply following a robot partner's instructions, participants in the Collaborative Team evaluation had to ask the robot questions, make decisions with the robot and rely on the robot for information. This increase in collaboration did not allow for the creation of an accurate model of workload in the human-robot teams.

Gathering information regarding the effects that working with a robot partner has on a human is crucial to developing accurate human performance models. Rather than creating a real-life simulation of each formation of a team or task assignment, an accurate performance model can offer information regarding a human's expected capabilities, performance, workload or affect given a set of input conditions. This research contributes human performance knowledge regarding a collaborative human-robot team.

4.5.5 Reaction Time Validation Evaluation Results

Response time to secondary task questions and Item Reaction Time were measured in the Collaborative Team evaluation Scenario. Secondary task questions included the On List response and

the Danger Level response. The time taken to respond to each question and the number of question repeats are presented in this section. Data from two participants in the H-R condition were omitted because they did not follow instructions for answering the questions.

4.5.5.1 Secondary Task Questions - Total On List Response Time

The mean Total On List Response Time was 2.18 (St. Dev. = 2.37) seconds (s) in the H-H condition and 5.72 (St. Dev. = 5.81) s in the H-R condition. A Kruskal Wallis test revealed a significant effect of condition on total response time to On List questions, $\chi^2(1) = 50.82$, $p < 0.001$. The H-R condition participants took a significantly longer time to respond to the questions than the H-H condition participants. Figure 20 depicts the mean response times by condition.

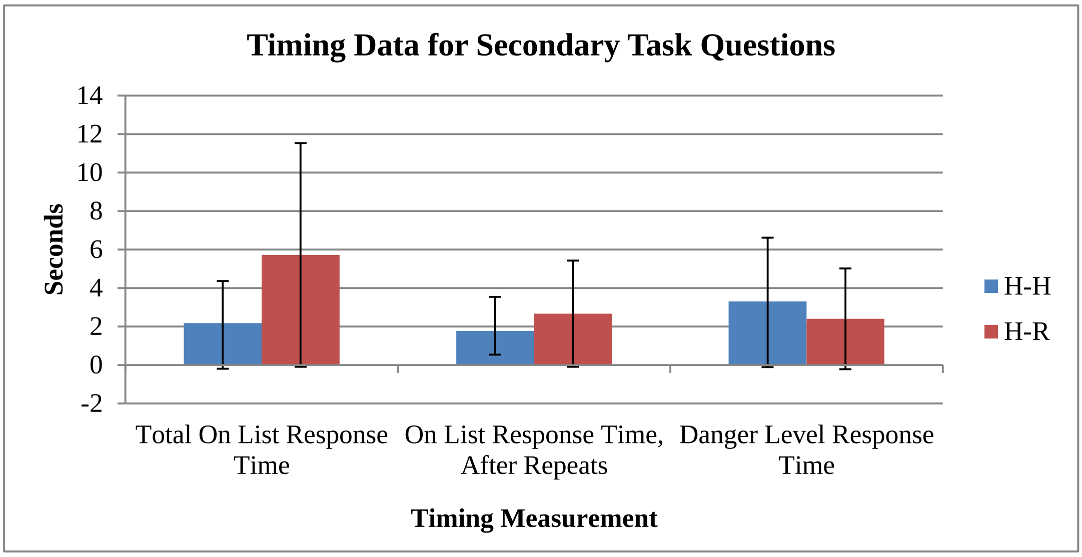


Figure 20. Timing Data for Secondary Task Questions, in seconds. Error bars represent one standard deviation above and below the mean.

The descriptive statistics for the Total On List Response Time by Investigation area and condition are presented in Table 15. A Kruskal Wallis test revealed a significant interaction effect between Investigation Index and condition, $\chi^2(5) = 56.95$, $p < 0.001$. A series of Mann-Whitney U tests revealed a set of significant differences, with a Bonferroni correction of $\alpha = 0.0033$.

The mean response time in Low Investigation Index areas independent of condition was 4.10 (St. Dev. = 5.02) s, 2.52 (St. Dev. = 3.14) s for Medium Investigation Index areas and 4.84 (St. Dev. = 5.26) s in the High Investigation Index areas. A Kruskal Wallis test revealed a significant effect of Investigation Index, $\chi^2(2) = 24.19$, $p < 0.001$. A series of Mann-Whitney U tests revealed significant differences between interactions of condition and Investigation Index, with a Bonferroni correction of $\alpha = 0.0167$. Low Investigation Index questions took significantly longer to answer than Medium Investigation Index questions, $U = 10311.5$, $z = 2.83$, $p = 0.002$. High Investigation Index questions also took significantly longer to answer than Medium Investigation Index questions, $U = 11516$, $z = 4.669$, $p < 0.001$. Questions answered in Low Investigation Index areas and High Investigation Index areas did not have significantly different response times. The Medium Investigation Index areas had a

shorter response time than the Low Investigation Areas. This discrepancy is due to a longer mean response time in Area 1 than in Areas 2 (Medium), 3 (Low) and 5 (Medium). Area 1's mean response time was 14.42 (St. Dev. = 26.12) s, while Area 3's (also a Low Investigation Index area) mean response time was 7.45 (St. Dev. = 11.20) s. Area 2's mean response time was 7.95 (St. Dev. = 12.62) s and Area 5's mean response time was 12.90 (St. Dev. = 17.62) s. The response time in Area 1 was high because the participants were learning to process the Secondary Task Questions and added to the list recall time. The High Investigation Index areas also had lower response times than Area 1. Area 4's mean response time was 10.34 (St. Dev, = 18.22) s and Area 6's mean response time was 11.42 (St. Dev. = 22.11) s.

The H-H condition resulted in a mean response time for the Medium Investigation Index areas that was significantly shorter than in High Investigation Index areas, $U = 3214$, $Z = 3.19$, $p = 0.001$. The H-R condition response times for Low Investigation Index areas were significantly longer than during Medium Investigation Index areas, $U = 2580$, $Z = 3.50$, $p < 0.001$. Response times in the H-R High Investigation Index areas were also significantly longer than in Medium Investigation Index areas, $U = 2611$, $Z = 3.86$, $p < 0.001$. H-R condition response times were significantly longer than the H-H condition response times in Low Investigation Index areas, $U = 3418.5$, $Z = 5.47$, $p < 0.001$, Medium Investigation Index areas, $U = 2616.5$, $Z = 2.78$, $p = 0.003$, and High Investigation Index areas, $U = 3142.5$, $Z = 4.04$, $p < 0.001$.

Table 15. Mean Total On List Response Times by Investigation Index and condition. Standard Deviations are provided in parentheses and times are reported in seconds.

Investigation Index	H-H	H-R
Low	1.97 (2.30)	6.44 (6.44)
Medium	1.73 (2.03)	3.44 (3.91)
High	2.82 (2.64)	7.13 (6.44)

Overall, H-R condition participants took a significantly longer time to answer questions regarding the presence of chemicals on the list, than the H-H condition participants. The area's Investigation Index played a role in the response times. The H-R condition participants had longer response times for all three Investigation Indices. The main effect of Investigation Index did not reflect that Investigation Index levels corresponded to the response times; Medium Investigation Index areas yielded the shortest response times, rather than the expected Low Investigation Index areas. One shortcoming of this analysis is that it does not account for the participant requested repeats of the questions or chemical names.

4.5.5.2 Secondary Task Questions – On List Response Time, after Repeats

The mean response time after removing repeats for the H-H condition was 1.77 (St. Dev. = 1.23) s and 2.67 (St. Dev. = 2.76) s for the H-R condition. A Kruskal Wallis test revealed no significant effect of condition on response time to On List questions when the timing began after the question was repeated. Figure 20 depicts the mean response times by condition.

The mean On List Response Time, After Repeats independent of condition was 2.34 (St. Dev. = 2.96) s for Low Investigation Index areas, 1.364 (St. Dev. = 0.674) s for areas with a Medium Investigation Index, and 2.58 (St. Dev. = 1.95) s for areas with a High Investigation Index. A Kruskal Wallis test revealed no significant effect of Investigation Index on response time to On List questions when timing began after the question was repeated. This test demonstrates that while it took participants longer to respond to questions in the Low Investigation Index, the difference is not significantly different from the Medium Investigation Index areas.

Mean response times to On List questions after repeats, separated by Investigation Index and condition are found in Table 16. A Kruskal Wallis test revealed no significant interaction effect of condition and Investigation Index on response time to On List questions when timing began after the question was repeated. Despite sizeable differences in the means between conditions, large standard deviations in the H-R data prevent significant differences.

Table 16. Mean On List Response Times, After Repeats by Investigation Index and condition.

Investigation Index	H-H	H-R
Low	1.57 (1.02)	3.15 (3.76)
Medium	1.50 (0.71)	1.33 (0.71)
High	1.94 (1.43)	3.04 (2.17)

Overall, once the time taken to repeat chemical names and questions were removed from the response time, there was no significant difference between conditions and no effect of Investigation Index.

4.5.5.3 Secondary Task Questions – Danger Level Response Time

Three negative response times occurred in both conditions. Three other "zero time" responses were completely removed from the H-H danger level response time data set and ten were removed from the H-R condition. The data points were removed because rather than answering an On List question with "yes" or "no," the participants in these thirteen cases responded instead with a number for the danger level, which is not the correct procedure for the secondary task questions. The mean danger level response time in the H-H condition was 3.31 (St. Dev. = 3.42) s and 2.40 (St. Dev. = 2.62) s in the H-R condition. A Kruskal Wallis test indicated no significant main effect of condition.

The mean Danger Level Response Time independent of condition for the Low Investigation areas was 2.98 (St. Dev. = 2.83) s, 2.83 (St. Dev. = 3.39) s in Medium Investigation Indexed areas and 2.59 (St. Dev. = 3.03) s in High Investigation Indexed areas. A Kruskal Wallis test indicated no significant main effect of Investigation Index.

Mean Danger Level Response Times are presented by Investigation Index and condition in Table 17. A Kruskal Wallis test indicated no significant interaction effect between Investigation Index and condition.

Table 17. Mean Danger Level Response Times by Investigation Index and condition.

Investigation Index	H-H	H-R
Low	3.59 (3.04)	2.28 (2.42)
Medium	3.50 (4.41)	2.25 (2.01)
High	2.24 (1.82)	2.90 (3.80)

Overall, Danger Level Response Time showed no significant differences by condition, Investigation Index or a combination of the two. This result indicates a similar length in response time between the two conditions, even though the results show longer response times in the H-H condition for the Low and Medium Investigation Index areas.

4.5.5.4 Secondary Task Questions – Number of Repeats

A total of 37 secondary task questions were repeated in the H-H condition and 69 were repeated during the H-R condition. The mean number of repeats per question in the H-H condition was 3.17 (St. Dev. = 3.30) and 5.75 (SD = 4.71) for the H-R condition. A Kruskal Wallis test revealed a significant effect of condition on the number of question repeats, $\chi^2(1) = 14.48$, $p < 0.001$, with the H-R condition participants requesting a significantly higher number of repeats. Figure 21 displays the total number of repeats for each question, by chemical name and condition.

The mean number of repeats independent of condition for questions in the Low Investigation Index areas was 5.25 (St. Dev. = 3.77). Questions in the Medium Investigation Index areas had a mean number of repeats of 1.50 (St. Dev. = 1.31) and there was a mean of 6.50 (St. Dev. =5.15) for the High Investigation Index areas. A Kruskal Wallis test revealed a significant effect of Investigation Index on number of repeats, $\chi^2(2) = 25.09$, $p < 0.001$). A series of Mann-Whitney U tests with Bonferroni-corrected α=0.0167 revealed that the number of repeats in the Low Investigation Index areas was significantly higher than the number of repeats in the Medium Investigation Index areas, U = 10274, Z = 2,77, p = 0.005. The number of repeats in the High Investigation Index areas was significantly higher than in Medium Investigation Index areas, U =10735, Z = 3.52, p < 0.001. The difference in the number of repeats in the Low and High Investigation Index areas was not significant.

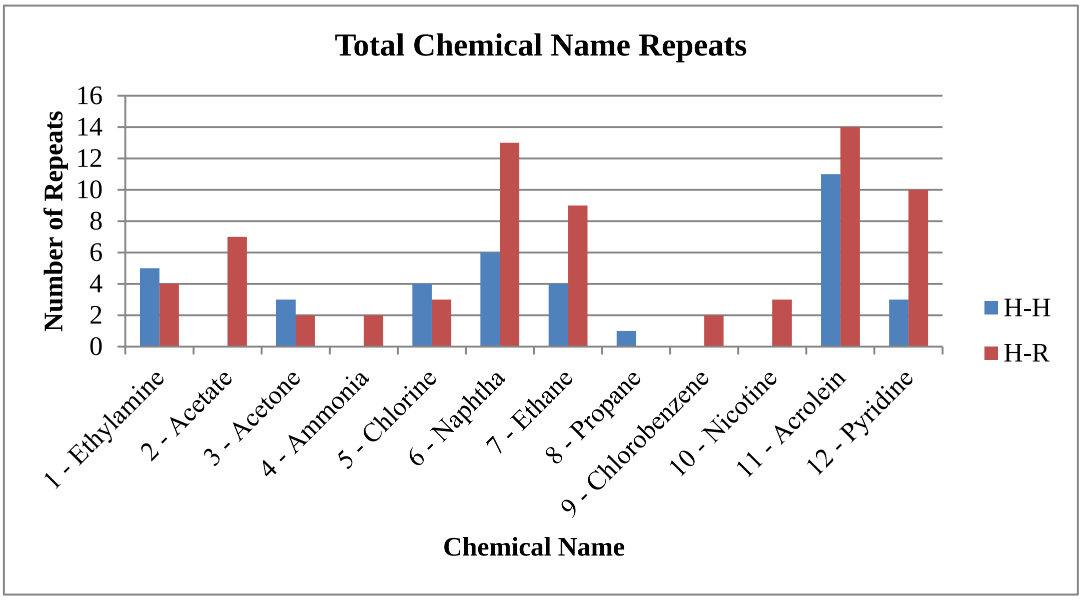


Figure 21. The total number of question repeats (y-axis) by chemical name in the secondary task question (x-axis) and condition.

The mean number of repeats by Investigation and condition are presented in Table 18. A Kruskal Wallis test revealed a significant interaction effect between Investigation Index and condition, $\chi^2(5) = 39.71$, $p < 0.001$. A series of Mann-Whitney U tests revealed a number of significant differences with a Bonferroni correction of $\alpha = 0.0033$. Specific to the H-H condition, the number of repeats in the Low Investigation Index areas was significantly higher than the number of repeats in the Medium Investigation areas, $U = 2008$, $Z = 3.099$, $p = 0.002$. Also, the number of repeats in the High Investigation Index areas was significantly higher than in the Medium Investigation Index areas, $U = 1907.5$, $Z = 3.72$, $p < 0.001$. The difference in the number of repeats between the Low and High Investigation Index areas was not significant. H-R condition participants had a significantly higher number of repeats in the Low Investigation Index areas than in the Medium Investigation areas, $U = 2321.5$, $Z = 2.19$, $p = 0.026$. The number of repeats in the High Investigation Index areas was significantly higher than in the Medium Investigation Index areas, $U = 2398.5$, $Z = 2.77$, $p = 0.005$. The difference in the number of repeats in the Low and High Investigation Index areas was not significant. There were no significant differences across the two conditions.

Table 18. Mean Number of Repeats by Investigation Index and condition.

Investigation Index	H-H	H-R
Low	3.75 (2.63)	6.75 (4.50)
Medium	0.75 (1.50)	2.25 (0.50)
High	4.75 (4.35)	8.25 (5.91)

Overall, the participants in the H-R condition asked the partner to repeat the question significantly more times. Both conditions' participants had a higher number of repeats in both Low and High Investigation Indexed areas, and a lower number of repeats in Medium Investigation Index areas. This significant difference affects the response time data, since the H-R condition participants significantly requested more repeats. Thus, response times were longer before the time for repeats was

removed. Once the larger number of repeats in the H-R condition was accounted for, there was no longer a significant difference between the response times to the On List questions.

4.5.5.5 Item Reaction Time

The mean Interaction Reaction Time in the H-H condition for the 26 experimenter-placed items was 5.97 s (St. Dev. = 13.68) and 15.32 (St. Dev. = 21.53) s for the H-R condition. Two outlier times were omitted. Some participants reacted to items resident in the environment that were not part of the evaluation, thus these item response times were excluded. A Kruskal-Wallis non-parametric test indicated that the H-R participants took significantly longer to respond to items, $\chi^2(1) = 50.18$, $p < 0.001$.

The mean Item Reaction Time for items, independent of condition, in Low Investigation Index areas was 10.74 (SD = 19.91) s, while the mean for the Medium Investigation Index areas was 11.36 (SD = 16.34) s and 10.74 (SD = 19.71) s for items in High Investigation Index areas. A Kruskal Wallis test found no significant effect of Investigation Index on reaction time.

Table 19 provides the mean Item Reaction Time for items by Investigation Index and condition. A Kruskal Wallis test revealed a significant interaction effect of Investigation Index and condition, $\chi^2(5) = 52.15$, $p < 0.001$. A series of Mann Whitney U tests revealed a set of significant differences with a Bonferroni corrected $\alpha = 0.0033$. Neither condition resulted in significant differences for item reaction times by Investigation Indices. The H-H condition participants had a significantly faster item reaction time than the H-R condition participants within the Low Investigation Index, $U = 1038.5$, $Z = -4.11$, $p < 0.001$, the Medium Investigation Index, $U = 1067.5$, $Z = -3.75$, $p < 0.001$ and the High Investigation Index, $U = 2266.5$, $Z = 4.56$, $p < 0.001$.

Table 19. Mean Item Reaction Time by Investigation Index and condition.

Investigation Index	**H-H**	**H-R**
Low	3.667 (3.772)	17.031 (25.630)
Medium	6.316 (9.124)	16.000 (19.864)
High	7.358 (19.485)	13.688 (19.537)

Overall, H-R condition participants took a longer time to respond to items present in their field of view. While there was an interaction effect between Investigation Index and condition, this result did not suggest that within each condition there were significantly different Item Reaction Times for the different Investigation Indices.

4.5.5.6 Item Find Type

Item Find Type was not represented in the Collaborative Team models. The most utilized Item Find Type in both conditions was a verbal utterance; participants indicated finding an item by verbally notify their partners more frequently than indicating a found object via touch, laser pointer, taking a picture or a visual fixation. The Item Find Type was only analyzed by condition, as further analysis by Investigation Index did not provide clear and useful insights. Table 20 provides the number of times

each Find Type was used when a participant began an item assessment. A Pearson's Chi-squared test indicated that there was a significant main effect of condition on Find Type, $\chi^2(4) = 31.00$, $p < 0.001$.

A series of Mann-Whitney U tests compared the number of found items by find type, in Table 20 with a Bonferroni correction of $\alpha = 0.0011$. There were no significant differences between the conditions for identical find types (e.g., H-H Touch vs. H-R Touch).

Considering comparisons within the H-H condition, the number of touch responses was significantly less than the number of verbal responses, U = 133, z = 4.35, $p < 0.001$, laser pointer responses, U = 170, z = 3.71, $p < 0.001$, and visual fixation responses U = 133, z = -4.35, $p < 0.001$. Total picture responses were significantly lower than the verbal, U = 155.5, z = 3.68, $p < 0.001$ and visual fixation responses, U = 159.5, z = -3.60, $p < 0.001$.

Comparisons within the H-R condition show that there were significantly more verbal responses than touch responses, U = 596, z = 5.20, $p < 0.001$, laser pointer responses, U = 587, z = -4.81, $p < 0.001$, and picture-taking responses, U = 605, z = -5.38, $p < 0.001$. There were also significantly more visual fixation responses than touch responses, U = 556, z = -5.38, $p < 0.001$, laser pointer responses, U = 587, z = -4.81, $p < 0.001$ and picture-taking responses, U = 605, z = -5.38, $p < 0.001$.

Overall, the most utilized Find Type in both conditions was a verbal utterance followed by a visual fixation. There were no significant differences between conditions for the number of finds in any one Find Type, indicating that participants did not respond to the items in significantly different ways.

Table 20. Number of times each Find Type was used in each condition.

Condition	Touch	Verbal	Laser Pointer	Took Picture	Visual Fixation
H-H	4	78	30	9	75
H-R	7	135	9	2	67

4.5.5.7 Comparing Modeled and Evaluation Results Reaction Time

The predicted On List response time was 1.51s for the H-H condition and 1.72 s for the H-R condition. On List response time, after repeats (see Section 4.5.5.2) was 1.77 (St. Dev. = 1.23) s in the H-H condition and 2.67 (St. Dev. = 2.76) s in the H-R condition. T-tests revealed no significant differences between either condition's modeled predictions and empirical results.

The predicted Danger Level response time was 1.44 s for the H-H condition and 1.65 for the H-R condition. Empirical results (see Section 4.5.5.3) showed a mean Danger Level response time of 3.31 (St. Dev. = 3.42) s in the H-H condition and 2.40 (St. Dev. = 2.62) in the H-R condition. T-tests revealed no significant differences between either condition's modeled predictions and empirical results.

The predicted Item Reaction Time was 1.41 s for both conditions. The mean Item Reaction Time (see Section 4.5.5.5) was 5.97 (St. Dev. = 13.68) s in the H-H condition and 15.32 (St .Dev. = 21.53) s in the H-R condition. A t-test indicated that there was no significant difference in the H-H condition. A t-

test in the H-R condition showed that the empirical Item Reaction Time results were significantly higher than the modeled prediction, $t(187) = 2.04$, $p = 0.04$.

Overall, the model was able to predict secondary task question response time for both questions and Item Reaction Time in the H-H condition. The model prediction of Item Reaction Time in the H-R condition however, was not a good predictor of the empirical results.

4.5.5.8 Reaction Time Results Discussion

The hypothesis was that response time is not affected by the participant's partner, human or robot. The results partially supported the hypothesis in that there were no significant differences in response times to secondary task questions (after repeat time was removed). Model predictions were also able to create a good prediction of secondary task question response time in both conditions and Item Reaction Time in the H-H condition. However, H-R condition participants took significantly longer than the H-H participants to indicate that they had found a potentially suspicious item and the model's predictions of this time were significantly lower than the empirical results.

The difference between the item response times is not negligible; the H-R condition participants required nearly three times the amount of time to respond to items in their field of view. However, the secondary task question response times indicate that the H-R condition participants were able to respond in a similar amount of time, thus the majority of H-R participants were mentally responsive. The question remains: why did participants in the H-R condition take longer to respond to items in the environment?

One hypothesis may be related to participants' assumptions regarding the robot's capabilities. One may argue that the participants assumed that the robot was on a set path throughout the evaluation areas and items to be assessed by the team were to be addressed in an order determined by the robot. The results, however, do not appear to support this hypothesis. Participants in both conditions did not significantly differ in the number of items that were identified. Thus, the participants were not simply waiting for the robot find each item or avoiding item assessment due to frustration with the robot's slow speech and movement.

A second hypothesis reflects a potential work under-load in the H-R condition, which can decrease signal detection. The mental workload reserve is reflected in responses to secondary task questions. The responses and response times were not significantly different across the conditions. The item response time is a task based more closely on vigilance principles. Signal detection rate decreases over time and is impacted by under-load and overload when monitoring the environment. A lower level of workload was found for the H-R teams. The addition of a robot teammate may lower workload enough for participants to potentially be in an under-loaded state. The result may be a greater effect of vigilance-related performance decrement than participants in the H-H condition. One avenue of future work will focus on the relationship between item saliency and response time, as more highly salient signals are less susceptible to vigilance decrements [66].

Yet another hypothesis is that the robot's slower and potentially difficult to understand speech required participants to allocate more attention to the robot, than the human partner, thus resulting in the participant focusing less attention on searching for items when the robot was speaking. The results do not appear to support this hypothesis. The H-H condition participants assessed an average of 24.11 (St. Dev, = 0.81) items, while the H-R condition participants assessed an average of 24.33 items (St. Dev. = 0.82), a difference that was not significant. A future analysis will investigate more closely what participants were doing during the evaluation, and where they placed their primary focus of attention. It is possible that the H-R participants were, for example, spending the majority of the evaluation focusing on the robot, while the H-H participants focused on investigating the environment.

The identification of reaction time trends in human-robot teams can be an important asset to the human-robot interaction community. One issue is a means of measuring reaction time in real-world evaluations. Typical reaction time evaluations require a participant to sit in front of a computer screen in a controlled setting and use metrics such as button presses to measure reaction time [51, 72, 73]. The presented research investigated methods for capturing reaction time in real-world settings. Using typical measurements of button press time or verbal utterance are not be suitable, because they are not natural interactions for the task. Laser pointers were given to participants to identify in the video at which items they were looking, but this method led to participants forgetting to identify the item with the laser pointer. It is necessary to get a signal from the participant to know when they have reacted, so simply mounting aerial view cameras to track participant movement cannot provide react time The ability to measure reaction time in such settings is critical to understanding and designing human-robot teaming capabilities. Tracking the Item Find Type helped identify the ways participants reacted to the items.

The underlying cause of the slower response time may be related to perceptions of robot capabilities and may be related to the participant's level of trust and willingness to rely on a robotic teammate. This evaluation included post-trial questionnaire results which may reflect a difference in team confidence, team leadership role, trust and understanding of partner capability. Further analysis of how the questionnaire data reflects the relationship between teammates is necessary. Means of further understanding this potential aspect of the results involve two approaches [33, 34, 39]. The first incorporates trust and reliance metrics into a similar evaluation with participants that have minimal training, while the second requires participants to training for extended periods with the robot prior to conducting the evaluation. It can be argued that extended training with the robot will allow the participants to have a better understanding and confidence in their robotic partner.

It is also possible that vigilance impacted the response time results. Maintaining vigilance is known to decrease task performance over time [46]. It is feasible that the presence of a robotic partner, in some manner negatively impacted the participants' vigilance during the evaluation. An analysis of a similar

task modeled in IMPRINT Pro demonstrated that the incorporation of vigilance into the model resulted in a 7.58% drop in the number of items found by the human, as presented in the next section.

5 HPMF Modeling

The primary focus of the validation efforts were the workload and reaction time HPMFs, however, we conducted modeling based research with a subset of the items in Table 1. Specifically the modeling activities focused on environmental aspects (cold/wind, heat/humidity and noise), personal protective gear (mission oriented protective posture –MOPP and Level A), Fatigue (sleepless hours), whole body vibration, and vigilance.

Specifically, IMPRINT Pro [1, 5] allows for the representation of the timing, workload and accuracy of tasks was used to conduct the model development. IMPRINT Pro provides micromodels of human behavior to help determine timing of tasks using established human factors data sets. For example, if a model contains the task for a human to walk 10 feet, the micromodels of behavior calculate the average time a person takes to walk that length. IMPRINT Pro also provides guidelines for assigning tasks' workload values, which combines values on seven workload channels: Auditory, Visual, Cognitive, Fine Motor, Gross Motor, Tactile and Speech workload. The values on each channel were assigned based upon channel guidelines. Using the example of walking 10 feet again, the Gross Motor workload value is based on walking on even terrain and there may be a visual component for looking where one are going or an auditory component for listening for directions, depending on the modeled situation. The accuracy of each task is determined by the modeler. The probability of success is the input. When the model executes, the task executes successfully based on the accuracy input. If the task fails, the modeler specifies what happens (i.e., a different task executes, the model ends or nothing happens).

Once a complete model is developed, different *stressors* can be applied to simulate different environmental and internal conditions that the human experiences. A stressor is a specific algorithm used to adjust expected performance under a particular set of conditions, which is similar to HPMFs [41]. For example, a model can be executed based upon specifying that that the human has not slept for 92 hours and it is 110 degrees Fahrenheit. These stressors impact the modeled human's performance. Modifying the stressors permits comparisons across model executions for different stressor values and combinations of stressors, such as the human having a full night of sleep and it is 70 degrees Fahrenheit.

5.1 Model Scenario

The test scenario models two team members (P1 and P2) investigating an area with each partner reporting any found suspicious items (this scenario is similar to the evaluation scenario in Section 4). If P1 finds something potentially suspicious, P1 calls over P2 to discuss the item and P1 reports if the item is suspicious. The roles are reversed if P2 locates a potentially suspicious item. The model

simulates the team conducting the search for 4 hours, and the analysis focuses strictly on P1's performance.

The team's search is divided into six search rounds lasting 40 minutes each. The maximum number of items that are modeled for a round is 27, thus the maximum number of total items 162. The items are interspersed throughout the simulated search with at least 100 seconds of search time between the identification of each item. Every fourth item is a suspicious item.

A high-level overview of the IMPRINT Pro model is provided in Figure 22. The tasks (purple rectangles with rounded corners) with P1 in parentheses represent P1's tasks, while those with P2 in parentheses are assigned to P2. P1 is concurrently searching, listening for communications from P2 and walking around the search environment. If an item is located, the team member stops the current task and enters the "(Px) Finds an item" function, where x represents the appropriate team member. The "(Px) Finds an item" is an example of a function that contains internal tasks, the gray rectangles with square corners in Figure 22. If "(P1) Finds an item" then the internal tasks are executed, the modeled tasks are provided in Figure 23.

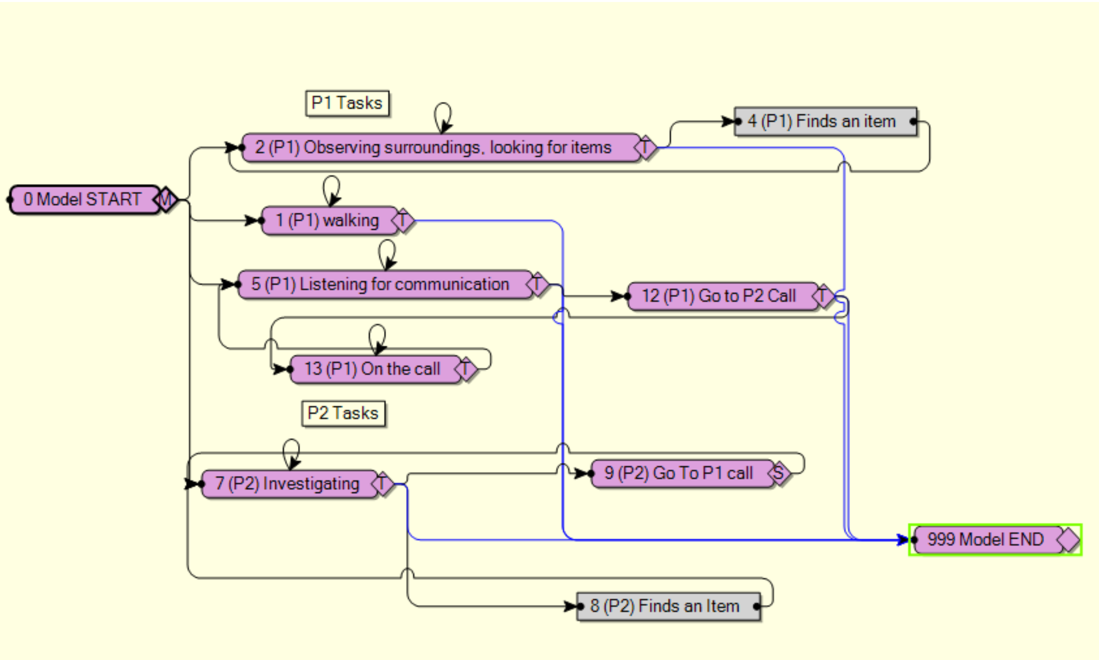


Figure 22. The high-level overview of model. Purple rectangles represent tasks and gray rectangles represent a bundle of tasks combined into a function.

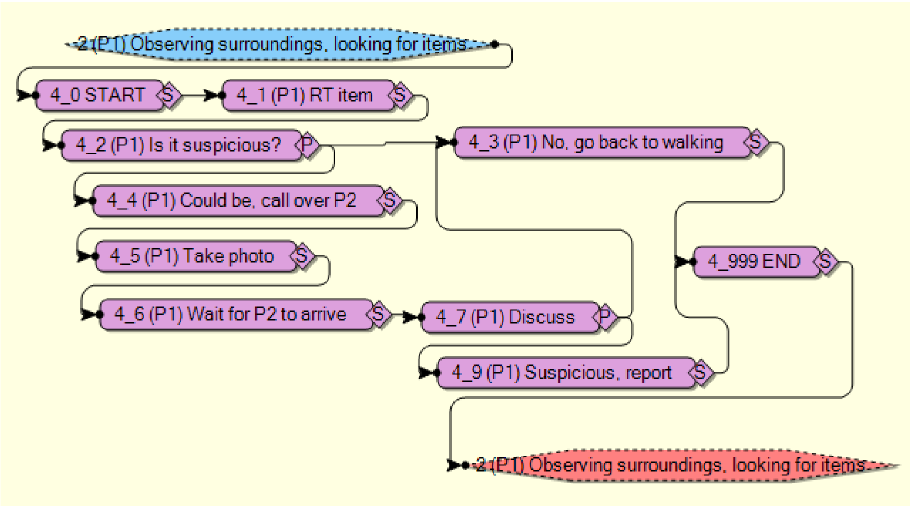


Figure 23. The internal tasks that compose the "(P1) Finds an item" function.

5.2 Modeled Moderator Functions

Stressors affect the timing and accuracy of task performance based upon weights associated with each task. A taxon is a way to categorize a task by its associated type of activity [41]. Nine taxons exist: Visual, Numerical, Cognitive (Problem Solving and Decision Making), Fine Motor Discrete, Fine Motor Continuous, Gross Motor Heavy, Gross Motor Light, Communication (Read and Write) and Communication (Oral). Each task can be assigned a weight for up to three taxons. Table 21 provides descriptions of each taxon, as defined in the Imprint Pro User Guide [41].

The IMPRINT Pro automatically calculates weights for each taxon within the model's tasks by using the modeler's provided workload values for the task. For example, tasks with high visual channel workload are weighted higher in the visual taxon. When applying the influence of a stressor on a series of tasks, the task timing and accuracy is affected by a) which taxons are affected by which stressors and b) the weights of the taxons in each task. For example, the Cold stressor affects the timing of tasks by impacting the weights of the Visual, Fine Motor Discrete and Gross Motor Light taxons. If a task does not provide a weight value for these taxons, then the task's timing is not affected by the Cold stressor. If the task is weighted highly in the taxons affected by the Cold stressor, then the task time increases.

Taxons impact task timing and accuracy. The task timing affects the number of items found by the searcher by increasing task length, making it more difficult to complete a higher number of task-finding tasks within a 40 minute search round. Stressors affect the task accuracy by making it more likely that the task execution fails. If a task crucial to finding an item fails, the model will not include that item on the list of found items. Two of the modeled tasks that have a negative impact on the accuracy of locating items are the "(P1) Is it Suspicious?" and the "(P1) RT Item". The first task represents the evaluation of an item as suspicious or not, while the second task represents the response time (RT) taken to notice the presence of a potentially suspicious item. The "(P1) Is it Suspicious?" task (Figure 23) has high weights in the Cognitive and Communication taxons. The "(P1) RT Item" task has high weights in the Visual and Cognitive taxons. Stressors that do not significantly increase

the weights of these particular taxons do not impact the total number of items found as much as other stressors.

Table 21. Taxon definitions and examples from the IMPRINT Pro User Guide [34].

Taxons	Definitions	Examples
Visual	Requires using the eyes to identify or separate targets or objects.	Seeing something move and recognizing it as an enemy tank.
Numerical	Requires performing arithmetical or mathematical calculations.	Measuring an azimuth on a map with a protractor.
Cognitive (Problem Solving and Decision Making)	Requires processing information mentally and reaching a conclusion.	Locating a fault in an electrical system after troubleshooting.
Fine Motor Discrete	Requires performing a set of distinct actions in a predetermined sequence mainly involving movement of the hands, arms, or feet with little physical effort.	Assembly and disassembly of the M-16 rifle.
Fine Motor Continuous	Requires uninterrupted performance of an action needed to keep a system on a desired path or in a specific location.	Tracking a moving target.
Gross Motor Heavy	Requires expending extensive physical effort or exertion to perform an action.	Loosening a very tight bolt with a wrench.
Gross Motor Light	Requires moving the entire body (that is, not just the hands) to perform an action without expending extensive physical effort.	Getting into a prone firing position.
Communication (Read and Write)	Requires either reading text or numbers that are written somewhere or writing text or numbers that can be read.	Reading a preventive maintenance check list for a vehicle.
Communication (Oral)	Requires either talking or listening to another person.	Giving a situation report by radio.

The modeled stressors include Cold and Wind, Heat and Humidity, Noise, MOPP gear, Level A gear, Sleepless Hours, Whole Body Vibration, Vigilance, and Response Time. Each stressor can be individually activated and set to any of the provided levels; when not in use, the stressor is set to a not applicable (NA) state and does not affect the model execution.

Cold and Wind – The Cold and Wind stressor incorporates a temperature range of -40 to 50 degrees Fahrenheit and wind speed from 0 to 50 knots. The time to complete a task increases based on the cold stressor, which affects the timing of Visual, Fine Motor Discrete and Gross Motor Light tasks.

Heat and Humidity– Heat and Humidity impacts negatively Visual, Numerical, Cognitive, Fine Motor Discrete, Communication (Read & Write) and Communication (Oral) task accuracies. The temperature range is 77 to 111 degrees Fahrenheit, while humidity ranges from 0 to 100%.

MOPP Gear – The mission oriented protective posture (MOPP) Gear is the protective clothing and equipment worn by military personnel. The five MOPP Gear levels range from Level 0 (the normal battle dress uniform with a carried mask and hood) to Level 4 (an overgarment and helmet cover, vinyl overboots, mask, hood and gloves). This stressor affects the timing of Visual, Fine Motor Discrete, Gross Motor Light and Communication (Oral) tasks.

Level A Gear – Level A gear is a suit that covers the entire body and head to protect the individual from hazardous situations and includes a breathing apparatus and gloves that severely limit hand dexterity. This stressor degrades both the timing and accuracy tasks in the Visual, Cognitive and Fine Motor Discrete taxons.

Sleepless Hours – The sleepless hours stressor incorporates how long the person has been awake and provides values ranging from 0 to 95 hours awake. This stressor negatively affects the timing and accuracy of Numerical and Cognitive tasks.

Whole Body Vibration – Whole body vibration is accounted for in terms of frequency and magnitude. The frequency refers to the amount of movement in the z-axis (see Figure 24), while the magnitude describes the severity of that movement (see Figure 25). This stressor degrades the accuracy of Visual, Fine Motor Discrete, Fine Motor Continuous and Communication (Read and Write) tasks. The timing of Fine Motor Discrete and Fine Motor Continuous tasks are increased. IMPRINT Pro currently only supports following frequency and magnitude combinations:

- High Frequency, Low Magnitude
- Medium Frequency, Low Magnitude
- High Frequency, Medium Magnitude

Vibration Frequecy Benchmarks:

Low (< 1 Hz) :	Motion sickness from ships, submarines, aircraft in moderate turbulence
Medium (1 - 4 Hz):	Vehicles at moderate speeds on uneven surfaces, aircraft in severe turbulence
High (> 4 Hz) :	Vehicles at high speeds on uneven surfaces, vessels at high speeds on choppy water

Figure 24. The Whole Body Vibration frequencies provided by IMPRINT Pro [41].

Vibration Magnitude Benchmarks:

Low :	Less than 0.63 m/sec²
Medium :	0.63 - 1.60 m/sec²
High :	Greater than 1.60 m/sec²

Figure 25. The Whole Body Vibration magnitude benchmarks provided by IMPRINT Pro [41].

Vigilance – Vigilance is the process of keeping an alert watchfulness [46] and a steady level of vigilance is difficult to maintain over time [87]. This stressor was developed by Vanderbilt and added to IMPRINT Pro as a time-based stressor, meaning that the impact of the stressor increases over time. Time-based stressors are added during modeling using data to support specific decrements to the task timing and accuracy. The literature demonstrates that signal detection decreases during a mission as time goes ensues [46, 54] and after two hours, signal detection decreases by approximately 28%, which was used for the presented stressor. The vigilance stressor negatively impacts accuracy of visual tasks, represented as a decrease in number of items found as time continues and the vigilance (looking for new items) was maintained.

Item Reaction Time – Item reaction time refers to the human searcher seeing an object, reacting to it and deciding to act on that item. The IMPRINT Pro provided micromodels for response time were

inaccurate, based upon data from the validation evaluation presented in Section 4.5. The IMPRINT Pro micromodels calculated a response time of approximately 1.41 seconds; however the validation determined that the mean item response time was 4.00 seconds (St. Dev. = 3.74 s), or approximately 2.84 times the modeled time. Thus the response time range was calculated using equation 3:

$$\textit{Response time} = 1.41 * 2.84 \pm 3.74 \tag{3}$$

Thus, the minimum response time was 0.2644 and the maximum was 7.744. This stressor affects the response time tasks by increasing the response time and the timing of the Visual and Cognitive (Problem Solving and Decision Making) taxons.

5.2.1 Results

The resulting model allows for an analysis of the impact that the various stressors on the human's performance. The base model results were derived from running the model with no stressors. A total of 20 trials were completed; however, the model trials were halted at the point where the simulation reached four hours of searching. The mean number of items found during the base model trials was 135.3 (St. Dev. = 2.8). Since the simulations were halted at four hours, the number of items found was lower than the total number of present items, 162. Performance decrements due to the application of each stressor in the following results are compared as a percentage decrease in items found from the number of items found in the base model condition: 135.3 items.

The model was analyzed for each individual stressor and a subset of stressor combinations. The average number of items found is calculated over the 20 trials.

5.2.1.1 Cold and wind

Pilcher et al.'s meta-analytic review of the effects of temperature on task performance found that tasks longer than one hour can result in a performance decrement of ~2.87% and when the entire task is longer than two hours performance can decrement by ~5.84% [58]. Cold and windy conditions affect task timing in IMPRINT Pro. The descriptive statistics for number of items found are presented in Table 22. The most extreme conditions, -40 to -22 degrees Fahrenheit with a 50 knot wind speed, yielded a 5.43% performance decrement when compared to the base model result. Overall, the average performance decrement was 4.35%, well within Pilcher et al.'s range for longer experimental times.

5.2.1.2 Heat and Humidity

Heat has been found to negatively impact performance. Specifically, Pilcher et al. [58] measured between a 2.67 and 2.71% performance decrement for heat. The accuracy of the modeled tasks is impacted by Heat and Humidity and the results are presented in

Table 23. The results indicate that heat and humidity impact the number of items found the most at temperatures above 94 degrees Fahrenheit and greater than 41% humidity. At higher temperatures and humidity, performance drops quickly; however, lower humidity does not have a large impact on

performance. The models predicted a 4.75% performance decrement from the base model for temperatures with a corresponding humidity below 51%. The high humidity levels lead to the identification of 22.25% fewer items than the No Stressors condition, while the harshest condition where P1 was able to find any items (>71% humidity and 103-111 degrees Fahrenheit) yielded a 78.30% decrease in in task accuracy. Once humidity in the 103-111 degrees Fahrenheit condition reaches over 81%, P1 is unable to find any items at all and the investigation task cannot be performed due to the extreme environmental conditions.

Table 22. Number of items found for the Cold and Wind stressor.

Stressor Conditions		Items Found	
Temperature	**Wind**	**Mean (St. Dev.)**	**Median**
-40F to -22F	**0 to 10 knots**	128.0 (2.6)	127.0
	> 50 knots	128.4 (3.7)	128.0
-21F to -4F	**0 to 10 knots**	127.8 (2.8)	127.5
	> 50 knots	130.1 (2.4)	130.0
-3F to 14F	**0 to 10 knots**	129.2 (2.9)	128.5
	> 50 knots	131.1 (3.4)	131.0
15F to 32F	**0 to 10 knots**	129.9 (3.1)	129.0
	> 50 knots	131.0 (4.2)	130.5
33F to 50F	**0 to 10 knots**	132.2 (3.1)	132.5
	> 50 knots	130.2 (3.0)	130.0

Table 23. Number of items found for the Heat and humidity stressor.

Stressor Conditions		Items Found		
Temperature	**Humidity**	**Mean**	**SD**	**Median**
85F to 93F	**0 to 10%**	132.9	3.2	133.0
	61 to 70%	132.9	3.2	133.0
	71 to 80%	131.2	3.8	131.0
	81 to 90%	131.2	3.8	131.0
	91 to 100%	131.2	3.8	131.0
94F to 102F	**0 to 10%**	132.7	3.2	132.0
	61 to 70%	129.7	3.8	129.0
	71 to 80%	125.0	4.4	126.0
	81 to 90%	118.8	5.4	120.0
	91 to 100%	108.3	3.6	108.5
103F to 111F	**0 to 10%**	132.7	3.2	132.0
	31 to 40%	129.0	3.4	129.0
	41 to 50%	117.2	2.9	117.5
	51 to 60%	111.3	3.9	111.5
	61 to 70%	47.1	4.6	47.0
	71 to 80%	29.4	5.9	29.0
	81 to 90%	0.0	0.0	0.0
	91 to 100%	0.0	0.0	0.0

5.2.1.3 MOPP Gear

Level 0-2 MOPP gear does not greatly affect the number of items found, as seen in Table 24. Level 3 MOPP Gear resulted in a 5.17% decrease in task accuracy, while the Level 4 MOPP Gear resulted in a 6.69% decrease in the number of items located.

Table 24. Number of items found for the MOPP Gear stressor.

Stressor Conditions	Items Found	
MOPP Gear Level	**Mean (St. Dev.)**	**Median**
Level 0	132.5 (2.8)	132.5
Level 1	131.5 (2.5)	131.0
Level 2	132.0 (2.1)	132.0
Level 3	128.3 (4.0)	128.0
Level 4	126.3 (2.3)	126.0

5.2.1.4 Level A Gear

The mean number of items found when simulating Level A gear was 70.6 items (St. Dev. = 3.9, median = 71). This result represents a 47.86% decrease from the base model results.

5.2.1.5 Sleepless Hours

The descriptive statistics for the number of items found based on the sleeplessness stressor are presented in Table 25. Little impact was found for sleepless hours of 0 to 24 hours; however, as the number of sleepless hours increased there was a large decrease in task accuracy. A 13.30% decline from the base model occurred with 25-47 hours, a 25.65% decline existed for 48 to 71 hours, and 72 to 95 hours awake declined the number of found items by 35.55%. Clearly, sleepless hours must be accounted for when predict a task performance accuracy.

Table 25. Number of items found for the Sleepless Hours stressor.

Stressor Conditions	Items Found	
Number of Sleepless Hours	**Mean (St. Dev.)**	**Median**
0 to 24 hours	131.9 (2.7)	132.0
25 to 47 hours	117.3 (3.5)	117.0
48 to 71 hours	100.6 (4.0)	101.0
72 to 95 hours	87.2 (2.7)	87.5

5.2.1.6 Whole Body Vibration

The IMPRINT Pro whole body vibration stressor is limited to three combinations of frequency and magnitude levels. Table 26 presents the number of found items. It is apparent that higher Frequency yielded a larger difference between the base model and this stressor. The most extreme condition (high frequency and medium magnitude) resulted in 14.38% fewer items being found than in the base model.

Table 26. Number of items found for the Whole Body Vibration stressor.

Stressor Conditions		Items Found	
Frequency	**Magnitude**	**Mean (St. Dev.)**	**Median**
Medium	**Low**	131.5 (3.6)	131.0
High	**Low**	122.0 (4.2)	122.0
	Medium	115.9 (3.4)	115.5

5.2.1.7 *Vigilance*

The addition of the time-based vigilance stressor helped to estimate the effect that searching for a prolonged period of time may have on vigilance. As time progresses, the visual accuracy decreases and it is less likely that all items will be found. An average of 125.1 (St. Dev. = 3.7, median = 125) items were found, which is 7.58% lower than the base model results.

5.2.1.8 *Adjusted Item Reaction Time*

The validation evaluation results reaction time resulted in slower recognition of out of place items, thus leaving less time for finding other items. When the adjusted reaction time was calculated using the mean empirical response time value, a mean of 129.9 (St. Dev. = 3.1, median = 129.5) items were found. This 3.99% decrement from the base model is a more realistic representation of how many items may be found in a real-life scenario. Alternatively, the number of items found was calculated using the mean empirical reaction time minus the standard deviation (Low End) and the mean empirical reaction time plus the standard deviation (High End). All results are presented in Table 27.

Table 27. Number of items found for the Adjusted Reaction Time stressor.

Stressor Conditions	Items Found	
	Mean (St. Dev.)	Median
Low End of Adjusted Response Time	134.1 (3.2)	134.0
Adjusted Response Time	129.9 (3.1)	129.5
High End of Adjusted Response Time	128.7 (3.8)	129.0
Unadjusted Response Time	133.1 (3.5)	134.0

5.2.1.9 *Multiple Stressors*

The modeling activity is also focusing on understanding how combining multiple stressors impacts the team's performance. IMPRINT Pro [41] provides five methods for conducting such an analysis. The reported results resulted from the Power Function approach that applied the inverse power to each task degradation in descending order of impact. As a result, the stressor with the most impact is given full effect.

A subset of the stressors was employed to demonstrate the impact on teaming. The first set of stressors included Sleepless Hours, Vigilance and Adjusted Reaction Time. Vigilance and Reaction Time were important additions to the analysis because the effects of these internal stressors are present regardless of the environmental conditions. Levels of Sleepless Hours showed a wide variation in items found in the single stressor analysis. The results, provided in Table 28, show a lower number of Items Found when compared to the stressors on their own (Vigilance – Section 5.2.1.7, Reaction Time – Section 5.2.1.8, and Sleepless Hours – Section 5.2.1.5).

The Heat and Humidity stressor was added to the analysis, whose results are also provided in Table 28. The analysis of only heat and humidity resulted in a large variation in Items Found for higher heat and humidity levels. Two combinations of heat (103-111 F) and humidity (41-50% and 61-79%) were added to the combined stressor analysis. The 41 to 50% humidity levels showed a large decrease in the number of Items Found from any of the stressors on their own. The results also resulted in a lower

number of Items Found when compared to the analysis combining Sleepless Hours, Vigilance and Adjusted Reaction Time. However, the 61-70% humidity level resulted in an extreme (93.3%) decrease in the number of items found when compared to the No Stressor condition.

Table 28. Number of Items Found in Power Function combined stressor situations, by stressor combination.

Stressor Conditions					Items Found
Vigilance	**Reaction Time**	**Sleepless Hours**	**Temperature**	**Humidity**	**Mean (St. Dev.)**
Vigilance Time Stressor	**Adjusted Response Time**	**25 to 47 hrs**	**N/A**	**N/A**	112.3 (4.9)
			103F to 111 F	**41 to 50%**	106.4 (4.2)
				61 to 70%	25.7 (7.7)
		72 to 95 hrs	**N/A**	**N/A**	86.8 (3.2)
			103F to 111 F	**41 to 50%**	85.4 (2.8)
				61 to 70%	9.1 (5.3)

Overall, the preliminary analysis of the number of Items Found by combining stressors showed that the number of Items Found was dramatically affected. Adding in the more realistic cases of multiple stressors can only add to the applicability and versatility of a human performance model and inform the design of human-robot interaction for such teams.

5.2.2 Discussion

It is impossible to represent every facet of a scenario within a human performance model, but the ability to incorporate the effects that the environmental state and the human's internal state may have on performance informs the task planning and assignment process (please see Section 6). Considerations of vigilance and response time, for example, should always be included in models involving signal detection in order to ensure a realistic detection rate. This analysis contributes a) the development and incorporation of new vigilance and response time stressors into IMPRINT Pro and b) the examination of the varying effects of different stressor conditions.

The analysis of the effects of environmental and internal stressors in IMPRINT Pro demonstrates the significant effect on performance that certain conditions have on task performance. The results from this analysis can be used to inform the allocation of team members to missions, by incorporating the performance predictions of individual team members towards successful completion of the overall mission objectives given specific conditions and circumstances. The three stressors that resulted in the largest change from the base model were: Level A gear, multiple nights without sleep and extreme humidity and heat conditions. While the impact of Cold, Vigilance and Response Time were not as dramatic, they add to the model's realism.

The allocation of team members to tasks and goals can be improved by accounting for the time and workload requirements necessary to complete a task. During some task deployments, environmental climate or necessary protective gear is unavoidable; however, understanding the impact of these constraints on task timing, accuracy and performance can influence the choice of teammates. Such predictions can lead to a higher probability of mission success, particularly in situations in which it is infeasible to repeat missions. Missions are costly, time consuming and labor intensive, thus the ability to quantify the potential mission outcome can be a critical tool. If it is known that a human will

encounter an extreme decline in performance accuracy given current conditions, the task can be divided into smaller sub-tasks, additional team members can be added or the tasks for the human and robot teammates can be allocated to ensure minimal impact on the human performance. Different combinations of task assignments can be analyzed via models before the mission success is put at risk.

Robots are not susceptible to decreases in performance due to internal stressors such as sleepless hours and may not be as vulnerable to potential hazards, thus avoiding performance decrements from stressors such as MOPP gear and Level A gear. Some tasks may be better performed by a human in an ideal set of conditions, but in others, swapping tasks between the human and robot teammates may be a better choice. Modeling tools, such as IMPRINT Pro that include built-in stressors and support modeling human and robotic entities provide a useful tool for understanding the human-robot teaming outcomes based on HPMFs. The results of such models, and results from validation evaluations can be integrated into autonomous allocation tools, such as Chazm (Section 6), that allocation missions to human-robot teams.

6 Chazm Model (CzM)

While OMACS was used in previous work to capture the basic entities required for task allocation among a set of artificial agents [19, 90], it is insufficient for use when considering humans as agents, especially when considering the effect of human performance on task execution. Thus, a new model was created that extended OMACS, called Chazm (CzM). CzM specifically captures human performance factors that can be used by task allocation algorithms. CzM provides a number of additions and modifications to OMACS to allow for capturing these human performance factors. Figure 26 shows some of the additions and changes to OMACS.

There are three types of elements in Figure 26: rectangles, lines, and ellipses. Rectangles are entities. Lines between entities are relations; the arrows on the lines denote direction for reading purposes only. For instance, role X requires capability *Y.* Ellipses attached to relations via a dashed line indicating values associated with relations that are functions. Elements that are greyed out are elements that exist in OMACS. In CzM, there are four new entities, six new relations (two of which are functions), and changes to existing elements.

6.1 Definitions

CzM defines an organization as a tuple $O = (G, R, A, C, P, F, Ch, T, \Phi$, achieves, requires, possesses, has, moderates, needs, uses, utilizes, contains, goodness) where

- G goals of the organization (same as the OMACS definition in Section 2.2)
- R set of roles (same as the OMACS definition in Section 2.2)
- A set of agents (same as the OMACS definition in Section 2.2)
- C set of capabilities (same as the OMACS definition in Section 2.2)
- P set of attributes
- F set of performance functions
- Ch set of characteristics
- T set of tasks
- Φ set of assignments, which are relations over $G \times R \times A$ (see Section 2.2)
- achieves relation over $\mathbb{P}(G) \times \mathbb{P}(R)$
- requires relation over $\mathbb{P}(R) \times \mathbb{P}(C)$ (same as the OMACS definition in Section 2.2)
- possesses function $A \times C \rightarrow [0, 1]$ (same as the OMACS definition in Section 2.2)
- has function $A \times P \rightarrow [-\infty, \infty]$
- moderates relation over $\mathbb{P}(F) \times P$
- needs relation over $\mathbb{P}(R) \times \mathbb{P}(P)$
- uses relation over $\mathbb{P}(R) \times \mathbb{P}(F)$
- utilizes relation over $\mathbb{P}(F) \times \mathbb{P}(F)$
- contains function $R \times Ch \rightarrow [-\infty, \infty]$
- goodness function $R \times A \times G \times \Phi \rightarrow [0, 1]$

Attributes. An *attribute* describes a property of an agent. Currently, there are three types of attributes: quality-type, quantity-type, and unbounded-type. A quality-type attribute constrains the value to the [0, 1] range, a quantity-type attribute constrains the range to [0, $+\infty$], and an unbounded-type attribute does not constrain the values [$-\infty$, $+\infty$]. In addition, each type is either a positive-type or negative-type attribute, which indicates the type of scale used to measure the values in relation to one another. Some attributes can be represented as either a positive-type (the higher the value, the better) or a

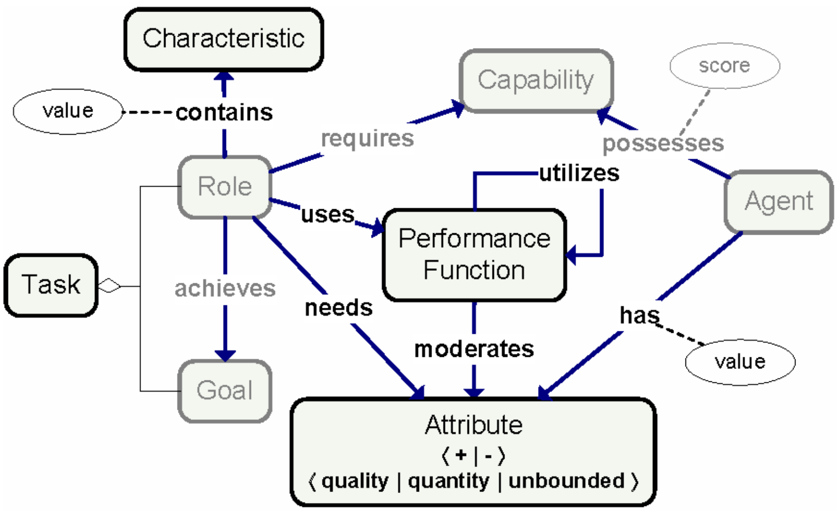


Figure 26. CzM Model

negative-type (the lower the value, the better). For example, consider the attributes *energy* and *fatigue*. These two attributes represent the same concept except that for energy, higher values are better, while for fatigue, lower values are better.

Performance Functions. The purpose of the second new entity, *performance function*, is to capture the HPMFs. Capturing HPMFs as an entity allows user-defined HPMFs to be used at runtime. For instance, two roles may have slightly different HPMFs for computing the *fatigue* of agents after performing different roles because one role may require more strenuous activities than the other. PMFs are captured in CzM as functions in the form of Definition 4.

$$\mathrm{pmf}_{\mathrm{attribute}} : \mathrm{R} \times \mathrm{A} \times \mathrm{G} \times \mathrm{Set}\{\mathrm{Assignment}\} \rightarrow [-\infty, \infty] \tag{4}$$

The *Role*, *Agent*, and *Goal* inputs inform the HPMF function to which role the agent is performing to achieve the goal. The *Set{Assignment}* is the relevant set of assignments for the HPMF function, which can be all the assignments of the organization or a subset such as the assignments of a particular agent; not all assignments affect the computation of HPMFs.

Characteristics. A *characteristic* describes a property of a role. A *characteristic* provides additional information that can be utilized by *performance functions*. For example, a role may contain information about the average length of time taken to complete the role, which can be captured as the *average completion time* characteristic. The *average completion time* characteristic can be used by *performance functions* associated with that role.

Task. A *task* is the composition of a role and a goal. The purpose of the *task* entity is purely for human understanding; computationally, a *task* does not provide any additional information other than what the associated role and goal already provide. In OMACS, an assignment is formally defined as *assignment*: $A \times R \times G$. In CzM, the definition of an assignment is expanded to include *assignment*: $A \times T$.

Has. The *has* function takes in an agent and an *attribute* and returns a value consistent with the type of that attribute: quantity $[0, +\infty]$, quality $[0, 1]$, or unbounded $[-\infty, +\infty]$. Even though the *has* function specifies a relation between an agent's *attribute* and a single value, it is straightforward to model complex attributes such as compound attributes. A compound attribute such as *location* does not contain a value but is comprised of multiple single values. For example, the *location* attribute is typically comprised of three values: longitude, latitude, and altitude. The three values can be represented as three attributes: *longitude*, *latitude*, and *altitude*. A logical grouping of the three attributes (*longitude*, *latitude*, and *altitude)* into the *location* attribute would not provide any functional benefits. To ease the use of CzM, design tools can provide logical groupings for complex attributes such as *location*. These design tools would then translate these complex attributes for use in CzM.

Moderates. The *moderates* relation specifies a relation between a *performance function* and an *attribute*. Because a *performance function* captures a HPMF and a HPMF computes the result for a

particular *attribute*, the *moderates* relation is a many-to-one relation (i.e., a *performance function* moderates exactly one *attribute* but an *attribute* can be moderated by multiple *performance functions)*. The *moderates* relation specifies the *attribute* to which the result of the HPMF is applicable. For example, to capture a HPMF that computes fatigue, the HPMF is captured as a *performance function* that *moderates* the *fatigue* attribute.

Needs. The *needs* relation specifies a relation between a role and an *attribute*. The purpose of the *needs* relation is to capture additional requirements for performing a role beyond just capabilities as currently used in OMACS. The *needs* and *requires* relations specify the complete set of requirements an agent must meet to perform a role.

Uses. The *uses* relation specifies a relation between a role and a *performance function*. The purpose of the *uses* relation is to indicate which of the *attributes* associated with the role through the *needs* relation require the use of a HPMF to compute the value. For example, the *reaction time* attribute may not need a HPMF because the value is obtained directly from the agent through the *has* function. But the *fatigue* attribute may need a HPMF to compute the value because the result may depend on the roles (e.g., *surveyor* role, *identifier* role, *rescuer* role). More importantly, the *uses* relation differentiates between attributes whose values are used and attributes whose values are changed as a result of performing roles. For correctness, there are two constraints on the *uses* relation: (1) a role can only use a performance function if the attribute modified by the performance function is also the attribute needed by the role (Constraint 5) and (2) a role cannot use two or more performance functions that moderate the same attribute (Constraint 6).

$$\forall r \in R, f \in F, a \in A \mid (r, f) \in \text{uses} \wedge (f, a) \in \text{moderates} \Rightarrow (r, a) \in \text{needs} \quad (5)$$

$$\forall r \in R, f, f' \in F, a \in A \mid (r, f), (r, f') \in \text{uses} \wedge (f, a), (f\ t, a) \in \text{moderates} \Rightarrow f = f' \quad (6)$$

Utilizes. The *utilizes* relation specifies a relation between two *performance functions*. The reason for the *utilizes* relation is to indicate whether a *performance function* uses another *performance function* for computation. For example, to compute the overall workload, the overall workload HPMF may require the auditory workload HPMF, cognitive workload HPMF, and visual workload HPMF. There are two constraints on the *utilizes* relation: (1) the *utilizes* relation do not form a cycle (Constraint 7) and (2) if a role uses a *performance function* (*A*), which utilizes another *performance function* (*B*), then the *attribute* moderated by *performance function* (*B*) is also needed by the role (Constraint 8). The transitive closure of a relation is denoted by the $^+$ symbol.

$$\forall f \in F \mid (f, f) /\in \text{utilizes+} \quad (7)$$

$$\begin{aligned}&\forall r \in R, f, f' \in F, a \in A \\ &\quad \mid (r, f) \in \text{uses} \wedge (f, f') \in \text{utilizes+} \wedge (f', a) \in \text{moderates} \Rightarrow (r, a) \in \text{needs}\end{aligned} \quad (8)$$

Contains. The *contains* function takes in a role and a *characteristic* and returns $[-\infty, \infty]$. For example, if a role takes 30 minutes to complete, that role *contains* the *average completion time* characteristic with a value of 30. Then a *performance function* for computing *fatigue* for that role can

use the *average completion time* characteristic for computing the new *fatigue* value of agents after performing that role.

Goodness. In OMACS, to perform a role, an agent must have the required capabilities. In CzM, to perform a role, an agent must have the required capabilities and the necessary attributes. The rcf function defined in OMACS is defined as *rcf* : *Role* × *Agent* → [0, 1] and the rcf function only evaluates the capabilities of an agent with respect to the role. This definition is no longer sufficient due to the addition of attributes; thus, the rcf function is not part of CzM. Instead, a goodness function is defined that evaluates both the capabilities and attributes required by agents. Furthermore, the specific goal being pursued is also part of the input for the goodness function because the goal may contain parameters that affect how well agents may perform a particular role. The *Set{Assignment}* is the relevant set of assignments for the goodness function, which can be all the assignments of the organization or a subset such as the assignments of a particular agent as not all assignments affect the computation of HPMFs. The goodness function has one constraint (Constraint 9), where the return value must be 0 if the agent does not possesses a required capability, a needed attribute, or the role cannot achieve the goal.

$$\text{goodness}(r,a,g,\phi) = \begin{cases} 0.0 & \text{if } \exists c \in (r,c) \in \text{requires} \mid (a,c) \notin \text{possesses} \\ & \quad \vee \exists n \in (r,n) \in \text{needs} \mid (a,n) \notin \text{has} \\ & \quad \vee (r,g) \notin \text{achieves} \\ [0,1] & \end{cases} \tag{9}$$

Achieves. In OMACS, the *achieves* function was defined as *achieves: R* × *G* → [0, 1], which indicates how well the role achieves a specific goal. However, in CzM, *achieves* is defined as a simple relation between a role and a goal. The reason for the change is because the goodness function takes in a goal as input, the score functionality is now part of the goodness function. This change provides greater flexibility because CzM allows specific goals to contribute to the goodness score. For example, if there are two agents capable of a task, sometimes it is preferable to select the agent that is physically closer to location where the task is to be completed.

6.2 Evaluation

Three scenarios were developed to evaluate the CzM model for the purpose of task allocations. The first two scenarios were created to evaluate the usefulness of the CzM model at runtime for making task allocation decisions and thus do not use validated HPMFs. However, the third scenario incorporates the HPMFs, as defined in Section 4 and the CzM model in a simulated human-robot team example.

6.2.1 Multiple Humans Evaluation

In the first scenario, the Conference Management System (CMS) [88, 18] is used as a basis for evaluating the usefulness of CzM versus OMACS. The CMS represents a conceptual model of the process that takes place leading up to a scientific conference, where authors submit their papers, reviewers are given papers to review, the PC chair makes decisions to accept papers, and accepted

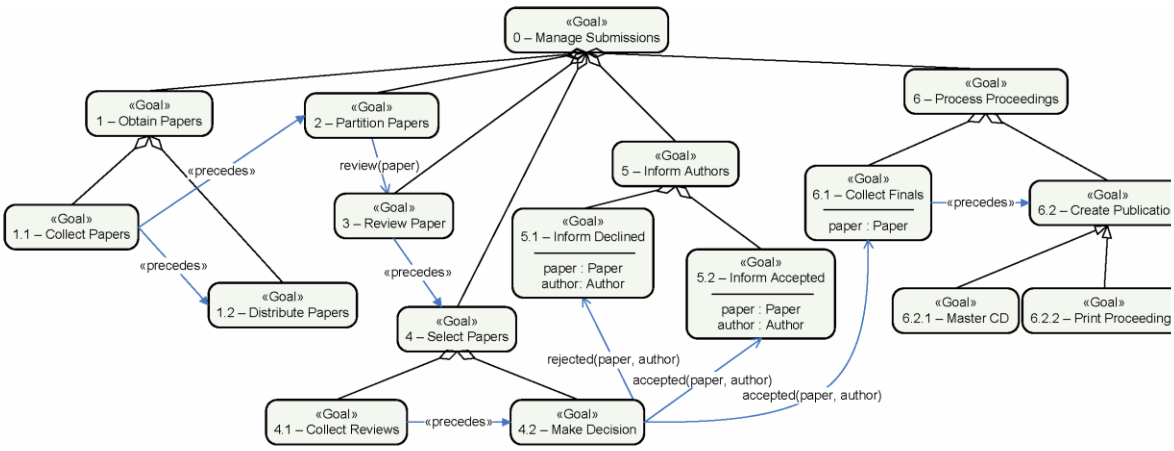


Figure 27. CMS Goal Model

papers are sent to the printers for printing. Figure 27 shows the Goal Model for Dynamic Systems (GMoDS) model that captures the CMS process, where system goals are represented and further decomposed into subgoals.

The top-level goal of the CMS is to *manage submi*ssions, which is decomposed into six conjunctive subgoals. These subgoals are also decomposed further into subgoals. At the bottom of the goal tree are the leaf goals, which are *collect papers*, *distribute papers*, *partition papers*, *review paper*, *collect reviews*, *make decision*, *inform declined*, *inform accepted*, *collect finals*, *master CD*, and *print proceedings*. These leaf goals are the only goals that can be pursued by agents to achieve the top-level goal.

The workload HPMF, which is the sum of the workload values of each paper the agent is reviewing, is defined by two equations: Equation 10 and Equation 11. In Equation 10, *p.type* refers to the type of paper. In Equation 11, *g.paper* refers to the paper parameter of the goal *g*.

$$\text{workload}(p) = \begin{cases} 10 & \text{if } p.\text{type} = \text{poster} \\ 20 & \text{if } p.\text{type} = \text{short} \\ 40 & \text{if } p.\text{type} = \text{full} \end{cases} \tag{10}$$

$$\begin{aligned} \text{pmf}_{\text{workload}}(r, a, g, \phi) = & \left(\textstyle\sum_{(a,r,g') \in \phi} \text{workload}(g'.\text{paper}) \right) \\ & + \text{workload}(g.\text{paper}) \end{aligned} \tag{11}$$

6.2.1.1 Experimental Setup

Each experiment assumes there are 50 reviewers that have to accept exactly 40 papers from a number of submitted papers, each of which has a quality in the range of [45%; 55%]. The range is kept small so as to increase the chance of a paper that is not in the top 40 being accepted due to inaccurate reviews of that paper; the small interval makes the problem harder because it is harder to discriminate between papers. Each submitted paper is given a random quality between [45%, 55%]. These submitted papers are ranked based on their quality; ideally, only the top 40 papers are accepted. There are a total of 80 experiments, which starts at 40 papers to review. The second at 41 papers to review,

the third at 42 papers to review, and so forth, up to the 80th experiment with 120 papers to review. Each submitted paper is reviewed by three reviewers. Once all reviews are in, the decision to accept or reject a paper is based on the three reviews. The purpose of an experiment is to evaluate how well a task allocation algorithm can assign these reviews to reviewers so that the set of accepted papers is as close as possible to the ideal set.

There are three types of reviewers defined: tenured professors, assistant professors, and graduate students. All reviewers need three attributes: *incentive*, *stress*, and *workload*. These three attributes are of the same scale, where a value of 0 means no incentive, no stress, and no workload respectively (determining what the values means in terms of numbers is beyond the scope of this research). *Incentive* values are none, low, medium, and high, maximum. For computational purposes, the incentive values are mapped to 0, 30, 50, 70, and 100 respectively. *Stress* and *workload* are measured in terms of percentages and do not have an upper-bound. These attributes determine the maximum number of papers a reviewer can review before becoming overloaded/overburdened. An overloaded reviewer will produce reviews that are less than 100% quality. As *incentive* increases, a reviewer is able to review more papers before becoming overburdened. As *stress* decreases, a reviewer is able to review more papers before becoming overburdened. Similarly, as *workload* decreases, a reviewer is able to review more papers before becoming overburdened. Table 29 shows the starting values of the attributes and the reviewing capability of the three types of reviewers, where a value of 1.0 means 100%. Since there are 50 reviewers in the experiments, there are 16 tenured professors, 16 assistant professors, and 18 graduate students.

Table 29: attribute and capability value of agent types for the multiple human evaluation

	Incentive	**Stress**	**Workload**	**Review Ability**
Tenured Professors	low (30)	0%	0%	1.0 (100%)
Assistant Professors	medium (50)	50%	0%	0.8 (80%)
Graduate Students	low (30)	60%	0%	0.6 (60%)

There are three types of papers defined: full paper, short paper, and poster paper. Reviewing a full paper would add 40% to a reviewer's workload; reviewing a short paper adds 20% to the workload; and reviewing a poster paper adds 10% to the workload. Furthermore, the maximum load is multiplied by the score of the agent's reviewing capability, which ranges [0, 1]. And thus, tenured professors have 100 max load, assistant professors have 80 max load, and graduate students have 60 max load. The quality (q_r) of a review produced by a reviewer is defined by Equation 12. For example, if a tenured professor has 6 short papers to review, the *workload* HPMF will return a result of 120% workload, which results in the quality of all 6 reviews being 100 ÷ (120 + 30 − 0) × 100 = 66.6%.

$$\text{max load} = 100 \times \text{reviewing capability}$$
$$\text{total load} = \text{workload} + \text{stress} - \text{incentive}$$
$$q_r = \begin{cases} 100 & \text{if total load} < \text{max load}, \\ \dfrac{\text{max load}}{\text{total load}} \times 100 & \end{cases} \tag{12}$$

Workload is computed based on the number of papers assigned to a reviewer, with each paper contributing either 10%, 20%, or 40% to the reviewer's workload. The quality of a review (q_r) and the quality of the paper (q_p) determines the review score (s) as defined in Equation 13. As the review quality (q_r) approaches to 0, the range of possible review scores approaches [0, 100]. For example, if $q_p = 60$ and $q_r = 80$, then $s = 60 + [-10, 10] = [50, 70]$.

$$s = \begin{cases} q_p & \text{if } q_r = 100, \\ q_p + \left[-\frac{100 - q_r}{2}, \frac{100 - q_r}{2}\right] & \end{cases} \tag{13}$$

Once all reviews are done, the average review score is computed for each paper since there are three reviews per paper. The papers are sorted by the average review score and the top 40 are accepted.

In each experiment, the performance of five reorganization algorithms is compared. The five algorithms are (1) *random*, (2) *round robin*, (3) *greedy*, (4) *attributes-greedy*, and (5) *attributes-enhanced*. The first three algorithms only use information available in the OMACS model, while the last two require information from the CzM model. Although the goal model captures the entire CMS process, the focus of the experiments is on allocating the instances of the *review paper* goal.

Random. For a given goal, the *random* algorithm randomly selects a capable agent and assigns that goal to the agent, where a *capable agent* is one that possesses all the required capabilities required to achieve a goal. This process continues until all goals have been assigned.

Round Robin. The *round robin* algorithm evenly distributes the goals to all capable agents. This process continues until all goals have been assigned.

Greedy. The greedy algorithm computes an aggregate score as shown in Equation 14 and assigns the agent with the highest score to the goal. This process continues until all goals have been assigned.

$$\frac{\text{goodness}(r, a, g)}{\text{number of papers assigned to } a} \tag{14}$$

The goodness function used is a simple average of all the required capability scores as shown in Equation 15.

$$\sqrt[|\{c|(r,c)\in\text{requires}\}|]{\prod_{c\in\{c|(r,c)\in\text{requires}\}} \text{possesses}(a, c)} \tag{15}$$

Attributes-Greedy. The *attributes-greedy* algorithm uses the goodness score to rank all agents for a given goal and assigns the agent with the highest score to that goal. This process continues until all goals have been assigned. Because the algorithm can access the workload, stress, and incentive values of an agent, the goodness function is defined as computing the review quality (q_r) as shown in Equation 15.

Attributes-Enhanced. The *attributes-enhanced* algorithm computes an aggregate score as shown in Equation 16. The agent with the highest score is assigned the goal. This process continues until all goals have been assigned. Because the algorithm can access the workload, stress, and incentive values of an agent, the goodness function is defined as computing the review quality (q_r) as shown in Equation 15.

$$\Delta_i = \mathbf{goodness} * (\text{assignments} + 1) - \Delta_{i-1} \tag{16}$$

The time complexity of the five algorithms are similar. If we assume that *g* is the number of unassigned goals, *a* the number of agents in the organization, *r* the number of roles in the organization, *c* the number of capabilities in the organization, and *n* the number of attributes in the organization. Then the time complexity of the random, round robin, and greedy algorithms is $O(g \times a \times r \times c)$, while the time complexity of the attributes-greedy and attributes-enhanced algorithms is $O(g \times a \times r \times (c+n))$. For a detailed proof, refer to [90]. Introducing attributes to CzM increases the time complexity of the goodness function by an expected amount.

Because of the randomness in various aspects of the experiments such as the random paper qualities and the bounded-random error for review scores, each experiment is executed 10,000 times to normalize the data.

6.2.1.2 Results

There are three types of data collected in the experiments: *score difference*, *set commonality*, and *review quality*. *Score difference* measures the sum of the accepted paper qualities versus the sum of the ideal paper qualities. *Set commonality* measures the percentage of ideal papers in the set of accepted papers. *Review quality* measures the average review quality of all reviews. In the experiments, there are two factors that significantly impact the performance of algorithms: the number of assignments[6] for each agent and the quality of reviews produced by each agent. The quality of reviews factor can only be measured accurately by algorithms with access to the workload, stress, and incentive attributes because these attributes affect the quality of a review as defined by Equation 15.

There is a relationship between the two factors; as the number of assignments increases, the quality of reviews tend to decrease. However, the importance of the two factors is not constant throughout the experiments. Because the number of reviewers are fixed at 50 for all experiments, the number of assignments is less important than quality of reviews when the number of submitted papers are low. However, as the number of submitted papers increases, the importance of the number of assignments also increases to a point where the number of assignments becomes more important than quality of reviews. Also, the importance of the two factors depends on the measurement system. For example, the quality of reviews factor plays a more important part in the *review quality* measurement than the other two measurements. Based on the relationship between the two factors, the hypothesis is that the results are generally split into three parts: (1) the first part is where the quality of reviews factor is the dominant factor while the number of assignments factor is minor, (2) the second part is where the two

factors are equally important, and (3) the third part is where the number of assignments factors is the dominant factor while the quality of reviews factor is minor.

The performance of the algorithms is linked to how the algorithms use the two factors. Although the random algorithm ignores both factors, indirectly and to a certain extent through random selection, the random algorithm uses the number of assignments factor. The round robin algorithm considers only the number of assignments factor and ignores the quality of reviews factor. The greedy algorithm considers the number of assignments factor and, in a limited degree, considers the quality of reviews by using the capability score while ignoring the attributes. The attributes-greedy algorithm considers only the quality of reviews factor and ignores the number of assignments factor. The attributes-enhanced algorithm considers both factors.

Figure 28 shows the *score difference* graph. The greedy and round robin algorithms drop off immediately at the beginning of the graph although the greedy algorithm maintains an advantage over the round robin algorithm. The advantage of the greedy algorithm over the round robin algorithm is due to use of the two factors. Although the greedy algorithm considers both factors, the quality of reviews factor is incorrect as the goodness function for the greedy algorithm ignores attributes, which results in poorer performance when compared to the attributes-greedy and attributes-enhanced algorithms. The attributes-greedy and attributes-enhanced algorithms still produce assignments that result in 100% quality reviews. At the first point of interest (around 58 submitted papers), the attributes-greedy and attributes-enhanced algorithms can no longer keep some reviewers from being overburdened. However, the attributes-greedy and attributes-enhanced algorithms still maintain an advantage over the greedy and round robin algorithms. At the second point of interest (around 70 submitted papers), the attributes-greedy algorithm begins to perform worse than the greedy algorithm probably because the number of assignments becomes a more important factor than the quality of reviews factor. The attributes-enhanced algorithm still maintains a small advantage over the other algorithms because it considers both factors. The round robin algorithm barely maintains an advantage over the random algorithm because it ignores the score of an agent's reviewing capability, which matters in these experiments. At the third point of interest (around 106 submitted papers), the performance of all algorithms seem to converge probably because situation is bad enough that any algorithm would perform just as well.

Figure 29 shows the *set commonality* graph. Again, the round robin and greedy algorithms drop off immediately at the beginning of the graph but the greedy algorithm, which considers both factors, maintains an advantage over the round robin algorithm. At the first point of interest (around 56 submitted papers), the attributes-greedy and attributes-enhanced algorithms are not able to keep some reviewers from being overburdened but they still maintain an advantage over the other algorithms. At the second point of interest (around 66 submitted papers), the greedy algorithm almost catches up to the attributes enhanced algorithm and the attributes-greedy algorithm begins to perform worse than the greedy algorithm. This change is due the number of assignments factor becoming the dominant

factor. At the third point of interest (around 104 submitted papers), the performance of all algorithms seem to converge probably because the situation is severe enough that any algorithm would perform just as good. Although, the attributes-enhanced algorithm seem to be slightly better the other algorithms.

Figure 30 shows the *review quality* graph. Again, the round robin and greedy algorithms start out worse than the attributes-greedy and attributes-enhanced algorithms. However, the greedy algorithm, which considers both factors, maintains an advantage over the round robin algorithm. At the first point of interest (around 58 submitted papers), the attributes-greedy and attributes-enhanced algorithms are no longer able to keep some reviewers from being overburdened but they still maintain an advantage over the other algorithms. At the second point of interest (around 68 submitted papers), the greedy

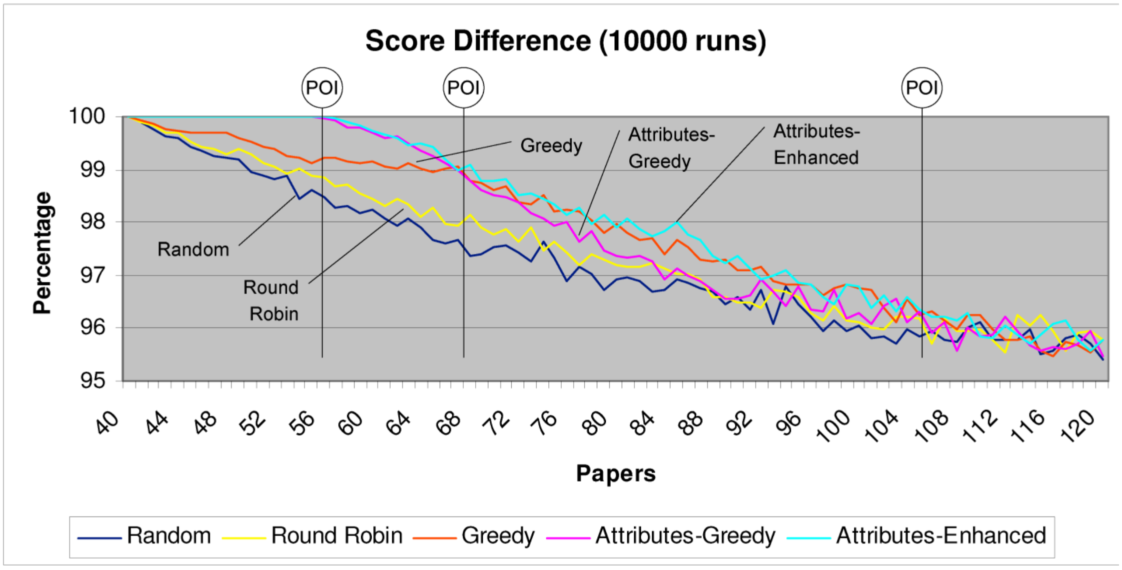


Figure 28. Score Difference Graph

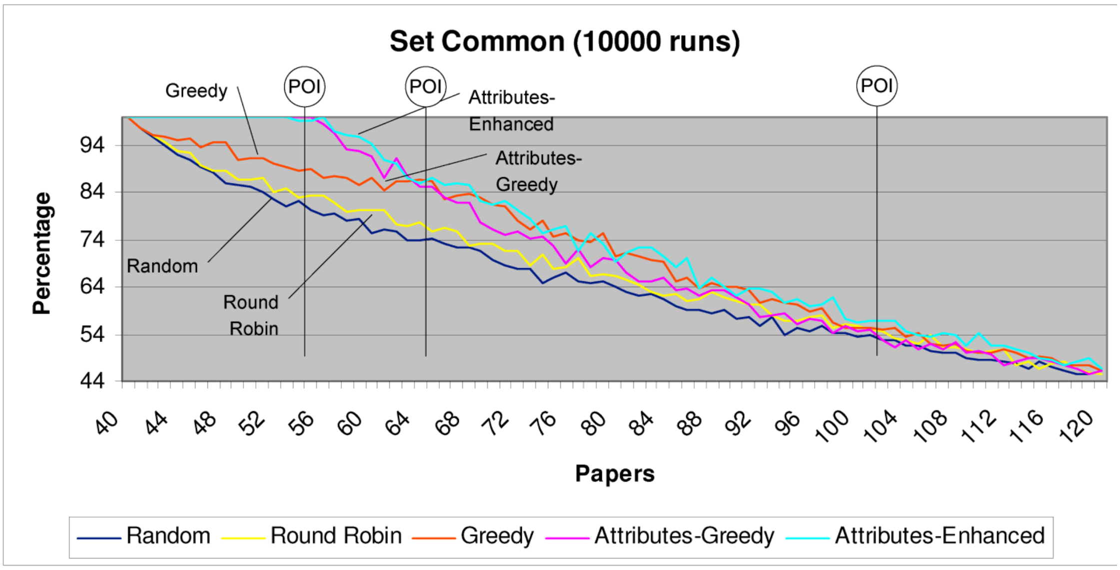


Figure 29. Set Commonality Graph

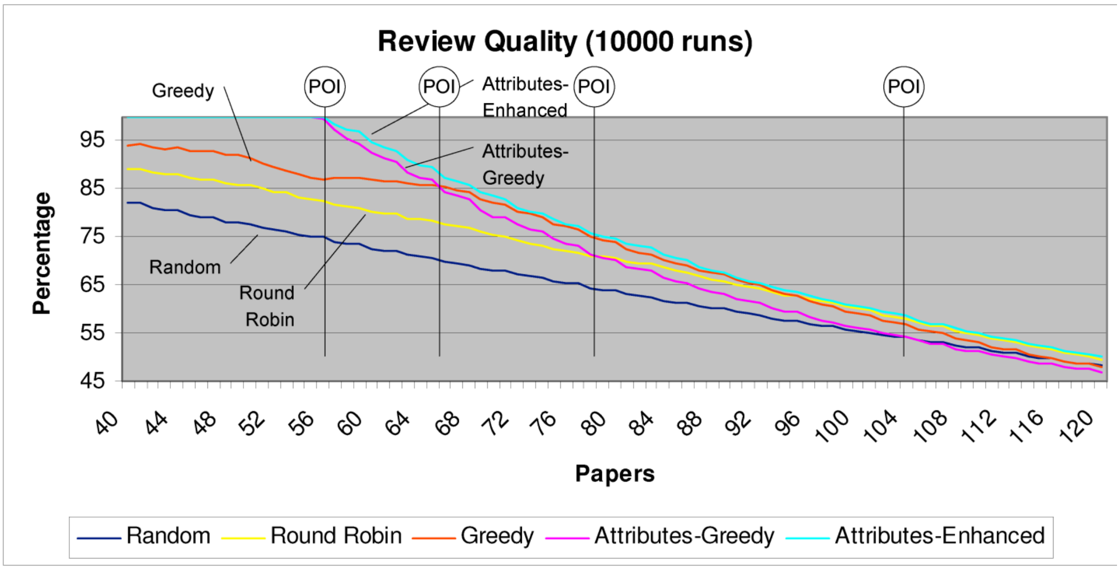


Figure 30. Review Quality Graph

algorithm surpasses the attributes-greedy algorithm because the number of assignments factor becomes just as important as the quality of review factor. The attributes-enhanced algorithm still maintains a slight advantage over the greedy algorithm because it considers both factors properly. At the third point of interest (around 80 submitted papers), the number of assignments factor becomes the dominant factor. This results in the attributes-greedy algorithm performing worse than the round robin algorithm. At the fourth point of interest (around 106 submitted papers), the attributes-greedy algorithm performs worse than the random algorithm. This is probably due to the overwhelming importance of the number of assignments factor over the quality of reviews factor. Also, the performance of the greedy algorithm is surpassed by the round robin algorithm probably because the greedy algorithm incorrectly considers the two factors. The attributes-enhanced algorithm maintains a slight advantage over the round robin algorithm because it considers the two factors properly.

With the introduction of attributes, algorithms that take advantage of this extra information are able to perform better. However, the caveat of this extra information is that it needs to be considered in the proper context as demonstrated by the attributes-greedy algorithm.

6.2.2 Multiple Humans Multiple Robots Simulation

The CMS scenario has three limitations: (1) all the agents are modeled as humans, (2) the bulk of the assignments occur at the same time (i.e., when deciding the papers to be assigned to reviewers), and (3) there is only one HPMF. The retrieval scenario addresses the limitations by having a mix of humans and robots, tasks that occur at different times, and multiple HPMFs.

The retrieval scenario has two recurring tasks: (1) retrieve an item and (2) relay messages. The retrieve task is the primary task where performance will vary based on the agent. The relay task is a secondary task that also affects the performance of the retrieve task. However, the performance of the relay task is constant regardless of the agent performing it. These tasks can occur at different times

and in different amounts. In addition, retrieve tasks can have different minimum completion time while the completion time of relay tasks is constant. An agent can only work on one retrieve task at a time but can work on multiple relay tasks at the same time. However, human performance degrades when a human is working on too many tasks at the same time. Furthermore, as humans continue to perform tasks, their performance will degrade over time.

Representing the two recurring tasks are two goals: *retrieve* and *relay*. Respectively, two roles are defined for the two goals: *retriever* and *relayer*. The *fatigue* and *workload* attributes are necessary to perform either role, while humans and robots are capable of performing either of the roles.

The workload HPMF computes the workload of a given agent, which is the sum of the workload of all the tasks that are currently assigned to the agent plus a task that may be assigned to the agent. The workload HPMF is defined by two equations: Equation 17 defines the workload of the retrieve and relay tasks (i.e., the retrieve and relay goals respectively) and Equation 18 defines the workload of a given agent based on its current set of assignments plus a new task.

$$\text{workload}(g) = \begin{cases} 55\% & \text{if } g = \text{retrieve} \\ 15\% & \text{if } g = \text{relay} \end{cases} \tag{17}$$

$$\text{pmf}_{\text{workload}}(r, a, g, \phi) = \begin{cases} 0 & \text{if } a = \mathit{robot} \\ \left(\sum_{(a,r',g')\in\phi} \text{workload}(g')\right) & \\ \quad +\text{workload}(g) & \end{cases} \tag{18}$$

The fatigue HPMF computes the fatigue of a given agent, which is the sum of the fatigue at the completion of all currently assigned tasks plus a task that may be assigned to the agent. The fatigue HPMF is defined by two equations: Equation 19 defines the fatigue at the completion of the retrieve and relay tasks and Equation 20 defines the fatigue of a given agent based on its current set of assignments plus a new task.

$$\text{fatigue}(g) = \begin{cases} 3\% \times g.\text{distance} & \text{if } g = \text{retrieve} \\ 2\% & \text{if } g = \text{relay} \end{cases} \tag{19}$$

$$\text{pmf}_{\text{fatigue}}(r, a, g, \phi) = \begin{cases} 0 & \text{if } a = \text{robot} \\ a.\text{fatigue} & \\ \quad + \left(\sum_{(a,r',g')\in\phi} \text{fatigue}(g')\right) & \\ \quad +\text{fatigue}(g) & \end{cases} \tag{20}$$

6.2.2.1 Experimental Setup

An experiment has (1) six agents, (2) ten relay tasks, (3) a given number of retrieval tasks, (4) a time range [1, 15] in which the recurring tasks can appear, and (5) a range [1, 10] for the distance parameter of the retrieval goal. There are a total of ten experiments. The first experiment starts at five retrieval tasks, the second at ten retrieval tasks, the third at fifteen retrieval tasks, and so forth, up to the tenth experiment with 50 retrieval tasks. The purpose of the experiment is to evaluate how well a

task allocation algorithm can assign the recurring tasks so that the time taken to complete all tasks is as short as possible.

In each experiment, there are two brute force algorithms. One for OMACS and one for CzM. Due to the brute force nature of the algorithms, the two algorithms are similar where the difference is in how the scores are computed. However, the time complexity of the two algorithms are exponential, which are $O(a^g \times (g \times c))$ and $O(a^g \times (g \times (c + n)))$; where g is the number of unassigned goals, a is the number of agents in the organization, c is the number of capabilities in the organization, and n is the number of attributes in the organization. This complexity necessitates the small number of agents and tasks. Refer to [90] for further details.

OMACS. The computation of the score for an assignment is defined by Equation 21. The computation of the overall score for a set of assignments is defined by Equation 22, where assigned(Φ) is the set of agents that are currently assigned, total(a, Φ) is the number of assignments for agent a, score(a) is the retrieval task score for agent a, relays(a, Φ) is the number of relay tasks assigned to agent a, and relays(Φ) is the number of all relay tasks.

$$\text{score} = \sqrt[|\{c|(r,c)\in\text{requires}\}|]{\prod_{c\in\{c|(r,c)\in\text{requires}\}} \text{possesses}(a,c)} \tag{21}$$

$$\sum_{a\in\text{assigned}(\Phi)} \left(\begin{array}{c} \left(\frac{1}{\text{total}(a,\Phi)}\right)\text{score}(a) \\ -\text{relays}(a,\Phi)\left(\frac{\text{relays}(a,\Phi)}{\text{relays}(\Phi)}\right)\left(\frac{\text{relays}(a,\Phi)}{\text{total}(a,\Phi)}\right) \end{array} \right) \tag{22}$$

CzM. The computation of the score for an assignment depends on the task type. The score for a *retrieval* task is defined by Equation 23 while the score for a *relay* task is defined by Equation 24. The computation of the overall score for a set of assignments is defined by Equation 25, where score(a, r, g) is the score for the assignment.

$$\begin{aligned} & d\Delta = \text{distance} * (1 - \text{ability}) \\ & f\Delta = \text{distance} * \text{pmf}_{\text{fatigue}}(r,a,g,\phi) \\ & w\Delta = \text{distance} * \max(\text{pmf}_{\text{workload}}(r,a,g,\phi) - 1, 0) \\ & \text{estimated completion time} = \text{distance} + d\Delta + f\Delta + w\Delta \\ & \text{score} = \frac{\text{distance}}{\text{estimated completion time}} + \text{distance} \end{aligned} \tag{23}$$

$$\begin{aligned} & f\Delta = \frac{1}{\text{pmf}_{\text{fatigue}}(r,a,g,\phi)+1} \\ & w\Delta = \min\left(\frac{1}{\text{pmf}_{\text{workload}}(r,a,g,\phi)}, 1\right) \\ & \text{score} = \frac{f\Delta + w\Delta}{2} \end{aligned} \tag{24}$$

$$\text{overall score} = \sum_{(a,r,g)\in\Phi} \text{score}(a,r,g) \tag{25}$$

There are three types of agents: capable humans, average humans, and robots. These types capture the differences in ability when performing the retrieval task; capable humans are the best, followed by

average humans, and finally the robots. However, the performance of human agents is affected by fatigue and workload, whereas the performance of the robots is consistent. As fatigue increases, human agents take longer to complete retrieval tasks. Similarly, as workload increases (beyond a threshold), human agents take longer to complete retrieval tasks. All three types of agents are equally capable of performing the relay tasks and the time taken to complete relay tasks is constant. Table 30 shows the starting values of the two attributes and the two capabilities for the three types of agents. Since there are six agents in the experiments, there are two capable humans, two average humans, and two robots.

Table 30. Attribute and capability values of agent types for multiple humans multiple robots evaluation

	Fatigue	Workload	Retrieval	Relay
Capable Humans	0%	0%	1.0 (100%)	1.0 (100%)
Average Humans	0%	0%	0.75 (75%)	1.0 (100%)
Robots	0%	0%	0.5 (50%)	1.0 (100%)

Performance for the retrieval task is based on fatigue, workload, ability, and the distance parameter of the retrieval task. As a baseline, a perfect agent (1.0 ability, 0% fatigue and workload, and is unaffected by fatigue and workload) would complete a retrieval task in *d* time units, where *d* = *g* distance. In general, Equation 26 defines the progress that an agent make in one time unit when performing the retrieval task.

$$\begin{aligned}
&d\Delta = \text{distance} * (1 - \text{ability}) \\
&f\Delta = \text{distance} * \text{fatigue} \\
&w\Delta = \text{distance} * \max(\text{workload} - 1, 0) \\
&\text{estimated completion time} = \text{distance} + d\Delta + f\Delta + w\Delta \\
&\text{progress} = \frac{1}{\text{estimated completion time}}
\end{aligned} \tag{26}$$

For example, if a capable human is given a retrieval task with a distance of 2. Then the progress for the first time unit is 1 because fatigue is 0 and workload is less than 1. However, in the next time unit, the fatigue of that agent has increased by 3% (Equation 19). So the progress for the second time unit is 1/2.06 and the fatigue of the agent increases by 2.06 × 2 distance × 3% fatigue ≈ 2.91%. But the total progress of the task is only ≈ 98.5% complete, so the agent takes a third time unit to complete the task, at which point the fatigue of the agent increased by ≈ 0.09% to a total increase of 6% fatigue.

Due to the randomness of the experiments (the time in which the two recurring tasks can appear and the distance for the retrieval tasks) and the exponential time complexity, each experiment is executed 1000 times to normalize the data.

6.2.2.2 Results

Two types of data are collected in the experiments: *cases* and *completion time*. *Cases* measured the number of runs in which the CzM algorithm performed worse than, equal to, or better than the OMACS algorithm. *Completion time* measured the average time in which a run took to complete all tasks.

Figure 31 shows the results of the two algorithms as measured by *cases*. When the number of retrieval tasks is low, there is a small number of cases where the OMACS algorithm performs better, a significant number of cases where the performance is the same, and a small number of cases where the CzM algorithm performs better. However, as the number of retrieval tasks increases the trend changes. When the number of retrieval tasks is over 20, in virtually every case, the CzM algorithm performs better than the OMACS algorithm.

Figure 32 shows the results of the two algorithms as measured by *completion time*. In addition to the results from the two algorithms, a third line is added to the graph. The third line is the completion time using only six perfect agents; perfect agents are unaffected by fatigue and workload and have the best ability when performing retrieval tasks. The reason for the third line is to provide an approximation of the lower bound for the completion time. Ideally, the third line should come from an optimal algorithm. However, due to the exponential time complexity, it would take too long to obtain the results. For example, on a small case (where there are 3 agents, 12 relay goals, and 2 retrieval goals),

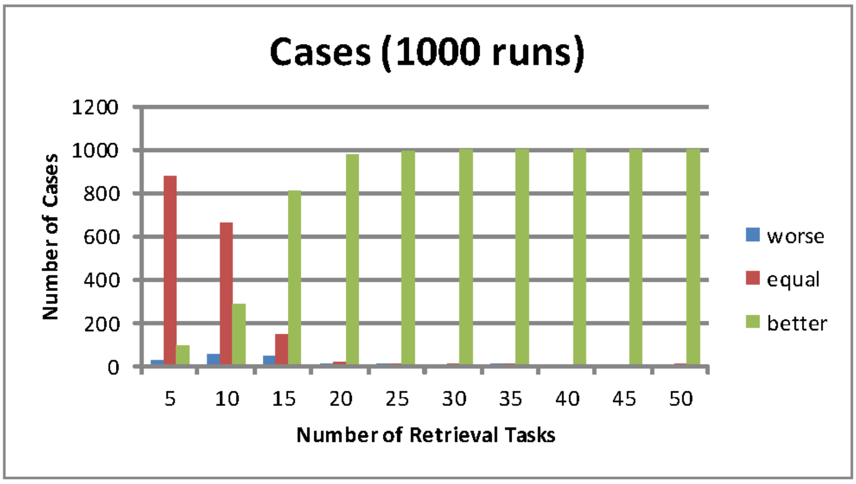


Figure 31. Cases Graph

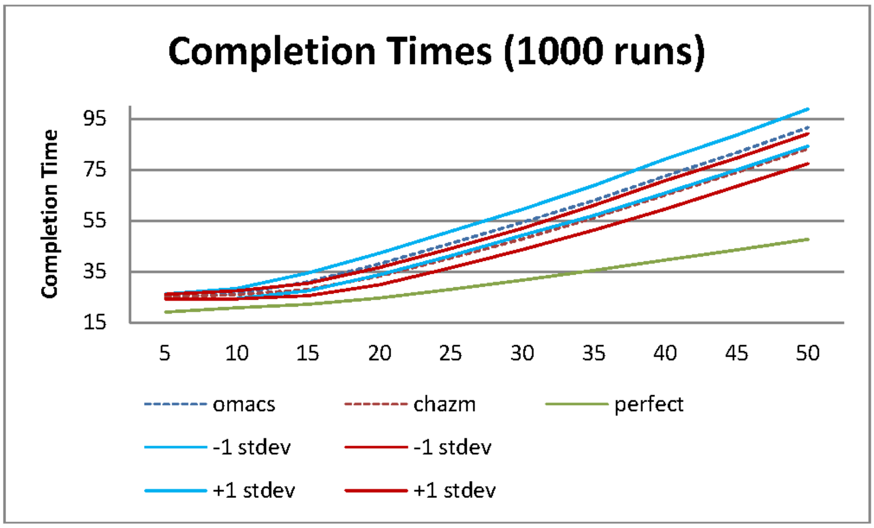


Figure 32. Completion Time Graph

there are approximately about 4 million (3^{12+2}) paths to explore per run. A path takes about 10 seconds (on an 8-core Intel Xeon E5462 at 2.80GHz with 12GB RAM) to explore, so exploring all 4 million paths would take about 1.2 years. However, the perfect agents line is a loose approximation because agents never tire or drop in their performance whereas the performance of human agents continue to deteriorate the more tasks they perform, more so towards the higher end (50 retrieval tasks). Even an optimal algorithm (with normal agents) cannot not perform better than the perfect agents line. Furthermore, the difference between the optimal line and the perfect line should be increasing as the number of retrieval tasks increases.

When the number of retrieval tasks is low, there is little difference between the two algorithms. However, as the number of retrieval tasks increases, the difference between the two algorithms becomes noticeable. The CzM algorithm maintains a noticeable difference (≈ 10% difference in terms of completion time) over the OMACS algorithm.

The results of this evaluation show that attributes and HPMFs in the CzM model can allow continuous task allocation algorithms to perform better when a mix of humans and robots are involved. The OMACS algorithm is already very good (within ≈ 80% of the perfect line at the early part of the results) and the CzM algorithm (an improvement of ≈ 10% over the OMACS algorithm) is also better in virtually every case when there are over 20 retrieval tasks.

6.2.3 Human-Robot Simulation Validation

The usefulness of human performance models at runtime was demonstrated by developing a simulation of the hazardous materials scenario. The simulation used the human performance model from Section 4 to calculate the effect of tasks on the human during scenario execution. While not complete, the purpose of the simulation was to demonstrate our envisioned use of human performance moderator functions in human-robot teams. The simulation is representative of the scenario described in Section 4 for Areas 1 and 2 and was created in the Cooperative Robotic Organization Simulator[4] (CROS). CROS is a multithreaded, grid-based environment for simulating multiagent systems designed around the Organizational Model for Adaptive Complex Systems (OMACS) [19]. CROS supports grid-based environments with a variety of object types and simulates the behavior of a set of heterogeneous agents within that environment. The simulation of the human-robot teams in CROS required that the OMACS model be extended to incorporate human performance moderator functions, as described in Section 6.1. A CROS simulation of the human-robot hazardous materials scenario was created along with scripted versions of the human and robot agents. The resulting simulated scenario was employed to validate that the human performance values computed by the human performance model during the simulation were consistent with those in the real experiments (Section 4).

4 See http://macr.cis.ksu.edu/cros

6.2.3.1 *Agent Simulator*

The scenario modeled in IMPRINT Pro was recreated in the CROS simulator to demonstrate the proposed approach for using human performance moderator functions in human-robot teams. Figure 33 shows a screenshot representing reconnaissance Areas 1 and 2 in the CROS environment. The dark brown lines represent the walls, while the various icons next to the walls represent the objects in those areas. The rectangle outlined by a dotted line in the upper right corner of Figure 33 is shown in more detail in Figure 34. All objects are positioned in the simulated environment as they were placed in the real world.

Two agents were simulated, one agent representing the human and another representing the robot. The robot (gray and white) and the human (black, blue and white) are shown in the middle of the hall in Figure 34, while a recycling bin (blue) and garbage bin (gray) are shown to the left of the agents along the wall and a bulletin board is shown below and to the right of the agents (brown). The yellow blocks

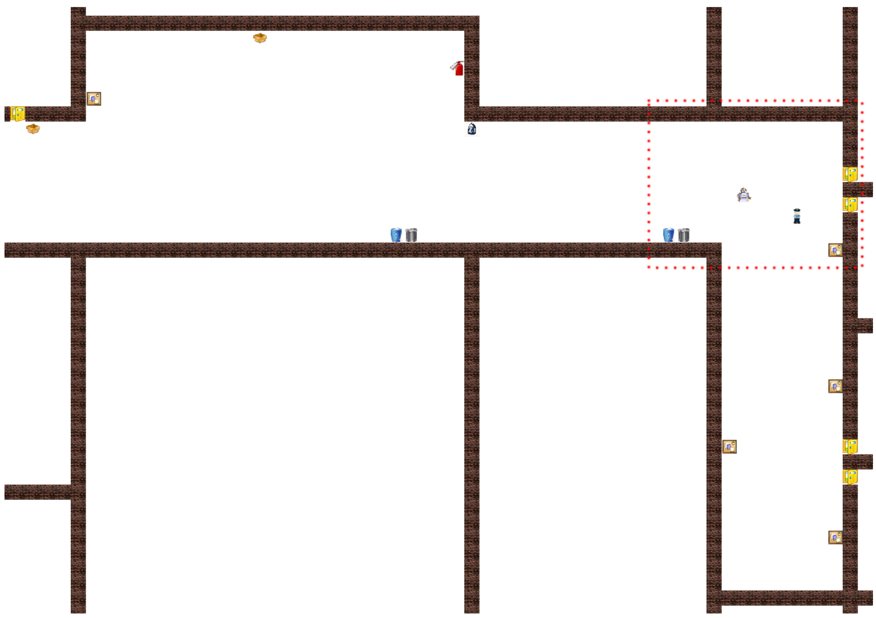


Figure 33. Simulated environment map for reconnaissance Area 1 and Area 2 that includes the robot, teammate, and various objects.

Figure 34. A zoomed view of the upper right corner of Figure 33, that better shows the robot, human, a recycling bin (blue), a garbage can (grey), and bulletin board (brown).

in line with the walls represent closed doors.

The human and robot agents possessed the same basic types of capabilities, including movement, localization, path planning, communication, vision, and air sampling. Although they had the same capability types, the actual capacities of those capabilities varied between human and robot. For example, the human could “see” further and higher than the robot. These differences in capabilities led to significantly different behavior in the simulation.

The first objective was to validate that the simulation produces similar workload results when compared to the real experiments. Thus, the initial simulation experiments defined the agent behavior in terms of a set of predefined scripts designed to match the actions from the real world experiment. As each task in the scenario was modeled separately and executed sequentially, it was straightforward to translate the modeled task behavior directly into simple scripted behavior.

Due to the large number of tasks modeled in the IMPRINT Pro (408 tasks in the six areas), the simulation used task categories to avoid the tedious process of creating a script for each individual task. The 408 tasks were categorized into 8 general categories: Walking, Listening, Speaking, Deciding, Reacting, Investigating, Taking Pictures, and Waiting. Scripts were created for each task type category.

The IMPRINT Pro model data for the scenario tasks was captured in six separate files, one for each reconnaissance area. An extra field was manually appended to the model data of each task to specify its category. A Java program was developed to convert the data files into script files that the simulator used directly. Essentially, the conversion program extracted five pieces of data from the files: the sequence number of the task, the task name, the task category, the total workload for the task, and the time taken to perform the task. Each task category, except Walking was extracted directly and reformatted into a new file. For example, the task to ‘Determine if it is necessary to go back to look at an unnoticed object’ from the scenario was converted into the following script.

```
Decide {
  index 19
  content Determine if going back is necessary
  workload 15.25
  timespan 3.0
}
```

The task category for this script is Decide, the sequence number (index) is 19, the name (content) is ‘Determine if going back is necessary’, the workload is 15.25, and the time taken to perform the task (timespan) is 3.0 seconds. The sequence number (index) is used to order the execution of tasks and synchronize the task executions between the human and the robot.

The Walk category required additional effort. The workload needs to be computed based on the time required to move from one location to another in the autonomous simulation, thus it was necessary to verify that the workload computed in the simulation (which is based on the time required to move the associated distance) was consistent with the workload from the IMPRINT Pro model. Since each grid

in CROS represents 1 foot (or approximately 0.305m) and the walking speed modeled in IMPRINT Pro is 1.612 m/s, we computed the average time an agent would spend in each grid while Walking at 0.189 seconds. Thus, by knowing the agent's start location and destination, the agent's path and the total time spent Walking can be calculated. Unfortunately, since CROS uses square grids, agents may only move horizontally or vertically, which dramatically increases the distance and time required to move from one location to another. Therefore, given the objective of validating the values from the real world experiment, the time calculations use the realistic assumption that humans tend to move in a straight line when possible, which is easily computed in a grid-based system using the Pythagorean Theorem. Thus, for the task 'Walk to BB1' (BB1 means Bulletin Board 1) the following script was generated, where the workload parameter is interpreted as the workload per second.

```
Walk {
  index 2
  content Walk to BB1
  workload 12.73
  timespan 2.28
  location (160, 71)
}
```

The scripts were parsed into the CROS simulator where each task was stored in a sequential list of *action task* objects. The human was simulated by executing each task in sequence from its list of action task objects. The robot was scripted by manually creating a similar set of data files based on the expected robot behavior from the real world experiments. Since workload is only related to the human agent, this information was omitted from the data for the robot. Scripting each agent in this manner enabled the simulated agents to perform the same basic tasks in the same sequence, as occurred in the real world experiments.

6.2.3.2 Validation of Simulated Results

All tasks in the scripted simulation, with the exception of Walking, are executed while the agents are stationary. Thus, the workload and time span values are used directly from the scripts, which were determined by the IMPRINT Pro model. This approach ensures that the IMPRINT Pro values match the values from the scripted simulation perfectly and no additional validation is required. However, since the Walk tasks required computing the time to move from one location to another, based on the actual simulator time, it was necessary to validate that, on average the times were consistent with those modeled in IMPRINT Pro.

Table 31 provides a comparison of the results of the Walk tasks between the scripted simulation and the values produced by the IMPRINT Pro model for Areas 1 and 2. The Workload Unit column represents the workload unit value assigned for this specific task in the IMPRINT Pro model. The Workload Units vary between tasks as the unit values are assigned to the Walking tasks based on the other tasks the human is doing. For instance, when walking to the bulletin boards, the human may have been looking at the bulletin board trying to ascertain what was on it, in addition to walking. The Script Time column represents the time calculated for the human to move to the next location, which

Table 31. Walk workload for Areas 1 and 2

Task Name	Workload Unit (w/s)	Scripted Time (s)	Model Time (s)	Scripted Workload (w)	Model Workload (w)
Walk to bulletin board 1	12.73	1.86	2.28	23.70	29.02
Walk to bulletin board 2	12.73	1.74	3.32	22.18	42.26
Walk to bulletin board 3	12.73	1.89	0.76	24.06	9.674
Walk to bulletin board 4	12.73	1.70	1.24	21.65	15.79
Walk to recycling bin 1	11.0	1.71	3.61	18.83	39.71
Walk to backpack	13.88	3.05	2.85	42.38	39.56
Walk to recycling bin 2	15.25	1.21	1.52	18.46	23.18
Walk to white board	15.25	4.25	2.66	64.77	40.57
Walk to box	14.11	1.97	1.9	27.84	26.81
Walk to book box	16.11	2.88	4.18	46.38	67.34
Walk to WL2	15.25	0.95	1.9	14.41	28.98
SUM		**23.21**	**26.22**	**324.66**	**362.90**
AVG				**13.99**	**13.84**

is based on an estimate of the distance in the grid-based environment. The Model Time column represents the modeled time in IMPRINT Pro. The Scripted and Model Workload columns represent the total workload for each task, as estimated in the simulation and IMPRINT Pro, respectively.

Obviously, the time spent on each Walk task does not match the IMPRINT Pro model exactly. This imprecision is caused by the fact that the distance is estimated based on a one foot grid size, which causes several location and distance errors. Notice, however, that when aggregated, the average walking workload in the scripted simulation is within two percent of the overall model workload. We believe this result is sufficiently close to support the investigation of using human performance moderator functions and the Chazm runtime model in human-robot team systems to predict human performance and to assign appropriate humans or robots to team roles in order to increase overall team performance.

6.2.4 Simulated Demonstration of CzM-HPMF Integration

The overall goal of this research was to show that we could use HPMFs in a human robot team to improve overall team performance by allowing the robot to better understand the human's capabilities and to allow it to adapt as the human's performance degraded over time. To demonstrate the usefulness of combining a validated HPMF with CzM for such situations, we extended the simulation presented in Section 6.2.3 to support estimating human workload in real-time based on the validated workload HPMF and assigning/re-assigning tasks from the human to the robot when the human's workload got too high.

To demonstrate this capability, we first eliminated the task scripts, described in Section 6.2.3.1, from both the robot and human agents in the simulation and replaced them with algorithms that implemented autonomous control of the robot and human actions. To do this, we essentially created an autonomous algorithm to carry out each type of task/goal (investigate, move, etc.) based on the parameters of the task (e.g., move to a *location*). A CzM model of the team was driven by the goal model shown in Figure 35. The scenario starts off with the *Supervise* goal, which is used to simply

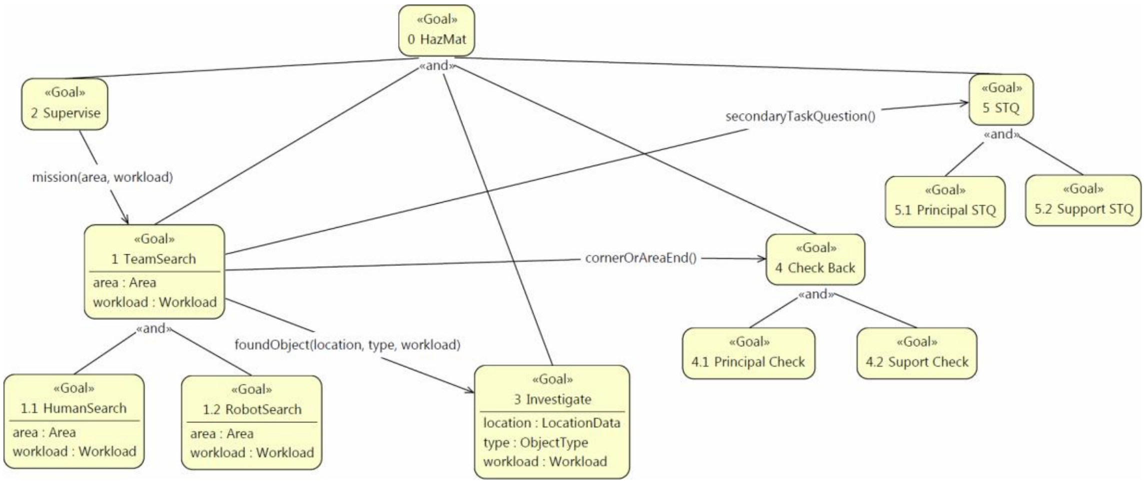


Figure 35. Autonomous Simulation Goal Model

initialize the simulation by triggering a *TeamSearch* goal that is decomposed into the *HumanSearch* and *RobotSearch* goals, which are assigned to the human and robot respectively. The *area* parameter is the area that the team should search, while the *workload* parameter specifies the limit for the human in terms of workload. This limit designates the point at which the robot should attempt to take over tasks normally assigned to the human. As the team performs the search, they can (1) find an object of interest, (2) reach a corner or end of a designated area, or (3) ask a secondary task question.

When an object of interest is found during the search, an *Investigate* goal is triggered with parameters describing its *location*, *type*, and the nominal *workload* associated with investigating the task. The *Investigate* goal is then normally assigned to either the human or the robot based on which one has the best capabilities (which in this scenario is always the human). The only time a robot would be assigned an *Investigate* goal is if the human's workload would surpass the *workload* limit. When the team reaches the end of an area or corner, a *Check Back* goal is triggered, which creates *Principal Check* and *Support Check* sub-goals, which are assigned to the human and robot and allows them to ensure the entire area has been searched before moving to the next area. Finally, when the human and robot need to communicate during their search or during an object investigation, the initiator of the conversations triggers a *STQ* (secondary task question) goal that allows the human and robot to converse. Triggering these goals allows the team to account for the human's workload.

To allow us to compare the effect of tracking the human's workload and using it in the assignment process, we first ran the autonomous agents without using the human workload in the assignment process. The results from this first experiment are shown in Figure 36 and Figure 37. Figure 36 shows the human's assignment for area 1 and 2 during the experiment. As shown, the human receives 11 different goal assignments, 1 *HumanSearch* and 10 *Investigate*. With these assignments, the human's overall workload is 102.5. Figure 37 shows the robot's assignment, which includes only the initial *RobotSearch* goal.

```
Human Total Assignments
<Human, HumanSearch, HumanSearch[1] (area=*x = 0, y = 0, width = 185, height=105, workload=*2.5)>
<Human, HumanInvestigate, Investigate[1] (location=(x=166,y=39), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[2] (location=(x=166,y=29), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[3] (location=(x=159,y=33), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[4] (location=(x=166,y=20), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[5] (location=(x=155,y=19), workload=10.0, type=Recyclingbin)>
<Human, HumanInvestigate, Investigate[6] (location=(x=142,y=12), workload=10.0, type=Backpack)>
<Human, HumanInvestigate, Investigate[7] (location=(x=137,y=19), workload=10.0, type=Recyclingbin)>
<Human, HumanInvestigate, Investigate[8] (location=(x=128,y=6), workload=10.0, type=Box)>
<Human, HumanInvestigate, Investigate[9] (location=(x=117,y=10), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[10] (location=(x=113,y=12), workload=10.0, type=Box)>
```

Figure 36. Human assignments without considering human workload performance

```
Robot Total Assignments
<Robot, RobotSearch, RobotSearch[1] (area=*x = 0, y = 0, width = 185, height=105, workload=*2.5)>
```

Figure 37. Robot assignments without considering human workload performance

Figure 38 and Figure 39 show the results of the same scenario except that during the scenario, the human's workload was taken into account when assignments were made. Figure 38 shows the human's overall goal assignments, which were reduced to 9: 1 *HumanSearch* and 8 *Investigate*. This reduction in goals occurred when new *Investigate* goals were triggered that, if assigned to the human, would have pushed the human's workload over the *workload* limit of 20. In both cases, the goals were assigned to, and carried out by, the robot as shown in Figure 39. With workload considered as part of the assignment process, the overall workload of the human was reduced by 20, from 102.5 to 80.5.

The result of this scenario only demonstrates that the workload HPMF can be used in conjunction with the CzM model to allow goals assignments to consider the human's performance based on their workload. As such, we have only demonstrated a framework for such decisions to be made; more research is needed to determine exactly *how* and *when* those decisions should be made.

```
Human Total Assignments
<Human, HumanSearch, HumanSearch[1] (area=*x = 0, y = 0, width = 185, height=105, workload=*2.5)>
<Human, HumanInvestigate, Investigate[1] (location=(x=166,y=39), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[2] (location=(x=166,y=29), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[4] (location=(x=166,y=20), workload=10.0, type=Board)>
<Human, HumanInvestigate, Investigate[5] (location=(x=155,y=19), workload=10.0, type=Recyclingbin)>
<Human, HumanInvestigate, Investigate[6] (location=(x=142,y=12), workload=10.0, type=Backpack)>
<Human, HumanInvestigate, Investigate[7] (location=(x=137,y=19), workload=10.0, type=Recyclingbin)>
<Human, HumanInvestigate, Investigate[8] (location=(x=128,y=6), workload=10.0, type=Box)>
<Human, HumanInvestigate, Investigate[10] (location=(x=113,y=12), workload=10.0, type=Box)>
```

Figure 38. Human assignments considering human workload performance

```
Robot Total Assignments
<Robot, RobotSearch, RobotSearch[1] (area=*x = 0, y = 0, width = 185, height=105, workload=*2.5)>
<Robot, RobotInvestigate, Investigate[3] (location=(x=159,y=33), workload=10.0, type=Board)>
<Robot, RobotInvestigate, Investigate[9] (location=(x=117,y=10), workload=10.0, type=Board)>
```

Figure 39. Robot assignments considering human workload performance

7 Conclusions

An issue that becomes apparent as multiagent systems become larger is that humans are no longer just users, humans are now peers working alongside agents, particularly robotic agents. A multiagent system's adaptability is increased because human peers can perform the work when the agents are unable to due to failures. However, there is a vast amount of information pertaining to humans that can be captured so that multiagent systems can work with their human counterparts. One of the first issues to be resolved is to decide what types of information about humans to capture. Often, human performance factors, which are often indicators of performance with respect to task performance, are the type of information that is first captured. These human performance factors are often captured as human performance moderator functions. Second, the information should be captured in such a way as to allow dynamic changes to occur because human performance factors often fluctuate over time.

Thus far, there has been very little focus on validating existing human performance moderator functions for application to peer-based human-robot teams. Twenty human performance moderator functions were modeled and two, workload and reaction time were validated for both human-human and human-robot teams. Workload was validated for two situations. The first required the human participant to follow their partner's instructions and provide information, while the second represented a collaborative relationship with specific responsibilities and a need for joint decision making tasks. The results indicate that a human-robot team results in lower workload for the human independent of teaming relationship; however, the reasons for the lower workload are not solely attributed to the slower navigation and speech speed of the robot. In fact, there are differing factors that appear to lower workload that dependent on the teaming relationship. Additionally, reaction time is significantly slower for human-robot partnerships, than for human-human teams. Reaction time is a difficult metric to capture in real-world environments; the methods of capturing this metric were identified and analyzed. Generally speaking, the modeling of differing human performance moderator functions across team compositions and relationships can guide the development of algorithms for allocating humans and robots to teams.

The Chazm model is able to capture information about the state of an agent through the *attribute* entity. This is the first step toward including humans as part of a multiagent system. The associated functions (*contains*, *affects*, and *needs*) provide the necessary structures so that multiagent systems can use this new information. In addition, the Chazm model can capture human performance moderator functions, which can be used for predicting human performance and for dynamically adjusting the associated values. The usefulness of the Chazm model for task allocation algorithms is demonstrated by the two scenarios, which show that the Chazm model can improve the results of task allocation algorithms. These validation results related to workload were used to simulate task allocation between a human and a robot team using the Chazm model.

References

1. Aasman, J., Mulder, G., and Mulder, L.J.M. Operator effort and the measurement of heart-rate variability. Human Factors **29** 161-170 (1987)

2. Allender, L. Modeling human performance: impacting system design, performance, and cost. Proc. of Military, Government and Aerospace Simulation Symposium, Advanced Simulation Technologies Conference. 139-144 (2000)

3. Allender, L., Kelley, T.D., Salvi, L., Lockett, J., Headley, D.B., Promisel, D., Mitchell,D., Richer, C., and Feng, T. Verification, Validation, and Accreditation of a Soldier-System Modeling Tool. Human Factors and Ergonomics Society Annual Meeting Proceedings **39** 1219–1223 (1995)

4. Anderson, J.R. and Lebiere, C. Atomic components of thought. Mahwah, NJ: Lawrence Erlbaum Associates (1998)

5. Archer, S., Gosakan, M., Shorter, P., and Locket III, J.F. New Capabilities of the Army's Maintenance Manpower Modeling Tool. Journal of the International Test and Evaluation Association **26**(1) 19–26 (2005)

6. Balch, T., and Arkin, R.C. Communication in Reactive Multiagent Robotic Systems. Autonomous Robots **1**(1), 27–52 (1994)

7. Benson, M., Koenig, K.L., and Schultz, C.H. Disaster triage: START then SAVE – a new method of dynamic triage for victims of a catastrophic earthquake. Prehospital and Disaster Medicine. **11**(2) 117-24 (1996)

8. "BioHarness Data Logger and Telemetry System." Data Acquisition **-** BIOPAC. Web. 31 Aug. 2010. <http://www.biopac.com/bioharness-data-logger-telemetry-system-acqknowledge>.

9. Blair, G., Bencomo, N., and France, R.B. Models@ run.time. Computer **42** 22–27 (2009)

10. Boles, D.B., Bursk, J.H., Phillips, J.B., and Perdelwitz, J.R. Predicting dual-task performance with the Multiple Resources Questionnaire (MRQ). Human Factors. **49**(1) 32–45 (2007)

11. Bomb Threat Plan – 'Code Amber.' UAMS Administrative Guide. Section 11.3.06. 09 Sept. 2008. Web 7 Sept. 2011 < http://www.uams.edu/adminguide/PDFs/11.3.06.pdf>

12. Botelho, S.C., and Alami, R. M+: a scheme for multi-robot cooperation through negotiated task allocation and achievement. IEEE International Conference on Robotics and Automation, **2** 1234–1239 (1999)

13. Bratman, M. Shared cooperative activity. The Philosophical Review, **101**(2) 327–341 (1992)

14. Carley, K.M., and Gasser, L. Computational Organization Theory. In G. Weiss (ed.) Multi-agent Systems: A Modern Approach to Distributed Artificial Intelligence, MIT Press, Cambridge, MA, USA (1999)

15. Choi, H.L., Brunet, L., and How, J.P. Consensus-Based Decentralized Auctions for Robust Task Allocation. IEEE Transactions on Robotics **25**(4) 912–926 (2009)

16. Dahla, T.S., Mataric, M., and Sukhatme, G.S. Multi-robot task allocation through vacancy chain scheduling. Robotics and Autonomous Systems **57**(6–7) 674–687 (2009)
17. Dash, R.K., Vytelingum, P., Rogers, A., David, E., and Jennings, N.R. Market-Based Task Allocation Mechanisms for Limited-Capacity Suppliers. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans **37**(3) 391–405 (2007)
18. DeLoach, S.A. Modeling Organizational Rules in the Multi-agent Systems Engineering Methodology. In: R. Cohen, B. Spencer (eds.) Advances in Artificial Intelligence: 15th Conference of the Canadian Society for Computational Studies of Intelligence, LNAI, 2338, 1–15. Springer-Verlag (2002)
19. DeLoach, S.A., Oyenan, W., and Matson, E.T. A Capabilities-based Model for Adaptive Organizations. Journal of Autonomous Agents and Multi-Agent Systems **16**(1) 13–56 (2008)
20. Dias, M.B. TraderBots: A New Paradigm for Robust and Efficient Multirobot Coordination in Dynamic Environments. Ph.D. thesis, Carnegie Mellon University (2004)
21. Dignum, V. A Model for Organizational Interaction: Based on Agents, Founded in Logic. Ph.D. thesis, Utrecht University (2004)
22. Dignum, V., Vazquez-Salceda, J., and Dignum, F. OMNI: Introducing Social Structure, Norms and Ontologies into Agent Organizations. In: PROMAS, 181–198 (2004)
23. "Disaster Preparedness & Response Network -Resources." Web. 15 Sep. 2010. <http://www.scahec.net/prepares/resources/media.html>.
24. Dorigo, M., Birattari, M., and Stutzle, T. Ant Colony Optimization. IEEE Computational Intelligence Magazine **1**(4), 28–39 (2006)
25. Foyle, D.C. and Hooey, B.L Human Performance Modeling in Aviation. Boca Raton: CRC Press (2008)
26. Fua, C.H., and Ge, S.S. COBOS: Cooperative Backoff Adaptive Scheme for Multirobot Task Allocation. In: IEEE Transactions on Robotics, **21**(6) 1168–1178 (2005)
27. "FR60." Garmin. Garmin Ltd. <https://buy.garmin.com/shop/shop.do?pID=27483>. Web. 7 Sept. 2011.
28. Gerkey, B.P., and Mataríc, M.J. Sold!: Auction Methods for Multirobot Coordination. In: IEEE Transactions on Robotics and Automation, **18(**5) 758–768 (2002)
29. Gerkey, B, Vaughan, R.T., and Howard, A. The Player/Stage project: Tools for multi-robot and distributed sensor systems. In Proceedings of the 11th International Conference on Advanced Robotics (ICAR 2003), 317-323 (2003),
30. Gerkey, B.P., and Mataric, M.J. A Formal Analysis and Taxonomy of Task Allocation in Multi-Robot Systems. The International Journal of Robotics Research, **23**(9), 939–954 (2004)
31. Gawron, V. J. Human Performance, Workload, and Situational Awareness Measures Handbook. 2nd ed. Boca Raton: CRC (2008).
32. Goodrich, M. and Schultz, A. Human-robot interaction: A survey. Foundations and Trends in Human-Computer Interactio*n*, **1** 203-275 (2007)

33. Groom, V., and Nass, C. Can robots be teammates? Interaction Studies. **8**(3), 483-500 (2007)

34. Hancock, P.A., Billings, D.R., Schaefer, K.E. A meta-analysis of factors affecting trust in human-robot interaction. Human Factors **53(**5), 517-527 (2001)

35. Hart, S. G. and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) *Human Mental Workload*. Amsterdam: North Holland Press (1988)

36. Harriott, C.E., Zhang, T., and Adams, J.A. Applying workload human performance moderator functions to peer-based human-robot teams. Vanderbilt University, Technical Report: HMT-2010-04 (2010)

37. Harriott, C.E., Zhang, T., and Adams, J.A. Evaluating the applicability of current models of workload to peer-based human-robot teams. In Proc. of 6th ACM/IEEE Inter. Conf. on Human-Robot Interaction,45-53 (2011)

38. Harriott, C. E., Zhang, T., and Adams, J. A. Predicting and Validating Workload in Human-Robot Teams. In Proceedings of the 20th Conference on Behavioral Representation in Modeling Simulation, 162-169, Sundance, Utah. (2011)

39. Hinds, P.J., Roberts, T.L., and Jones, H. Whose job is it anyway? A study of human-robot interaction in a collaborative task. Human-Computer Interaction. **19** 151-181 (2004)

40. Humphrey, C.M. and Adams, J.A. Robotic Tasks for CBRNE Incident Response. Advanced Robotics, **23** 1217-1232 (2009)

41. IMPRINT Pro User Guide: Volumes 1-3. (2009) *Alion Science and Technology*. <http://www.arl.army.mil/www/pages/446/IMPRINTPro_vol1.pdf> Web. 25 Aug 2012

42. Kephart, J.O., and Chess, D.M. The Vision of Autonomic Computing. Computer **36**(1) 41–50 (2003)

43. Kube, C.R., and Zhang, H. Collective Robotics: From Social Insects to Robots. Adaptive Behavior **2**(2) 189–218 (1993)

44. Kubo, M., and Kakazu, Y. Learning Coordinated Motions in a Competition for Food between Ant Colonies. Proceedings of the Third International Conference on Simulation of Adpative Behavior: From Animals to Animats **3** 487–492 (1994)

45. Macarthur, K.S., Stranders, R., Ramchurn, S.D., and Jennings, N.R. A Distributed Anytime Algorithm for Dynamic Task Allocation in Multi-Agent Systems. Proceedings of the Twenty-Fifth Conference on Artificial Intelligence. AAAI Press (2011)

46. Mackworth, N. H. The breakdown of vigilance during prolonged visual search. Quarterly Journal of Experimental Psychology, **1** 6–21 (1948)

47. Mahoney, P. F. Businesses and bombs: Preplanning and response. Facilities. **12**(10) 14-21 (1994)

48. Mataríc, M.J. Designing Emergent Behaviors: From Local Interactions to Collective Intelligence. Proceedings of the Third International Conference on Simulation of Adpative Behavior: From Animals to Animats 2, 432–441. MIT Press, Cambridge, MA, USA (1993)

49. McLurkin, J., and Smith, J. Distributed Algorithms for Dispersion in Indoor Environments Using a Swarm of Autonomous Mobile Robots. In: R. Alami, R. Chatila, H. Asama (eds.) 7th International Symposium on Distributed Autonomous Robotic Systems. Springer (2004)
50. Montemerlo, M., Roy, N., and Thrun, S. Perspectives on standardization in mobile robot programming: The Carnegie Mellon navigation (CARMEN) toolkit. Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2410-2414 (2003).
51. Nagler, C.A., and Nagler, W.M. Reaction time measurements. Forensic Science. **2** 261-274 (1973)
52. "NASA TLX Homepage." Human Systems Integration Division at NASA Ames. <http://humansystems.arc.nasa.gov/groups/TLX/>. Web. 15 Sept. 2010.
53. NATO. Civil Emergency Planning Civil Protection Committee. Guidelines for First Response to a CBRN Incident. <http://www.nato.int/cps/en/natolive/topics_49158.htm?selectedLocale=en>
54. Parasuraman, R. Vigilance, Monitoring, and Search. Handbook of Perception and Human Performance Vol. 2 Cognitive Processes and Performance. New York, NY: John Wiley & Sons, (1986)
55. Parker, L.E. ALLIANCE: An Architecture for Fault Tolerant Multirobot Cooperation. IEEE Transactions on Robotics and Automation **14**(2) 220–240 (1998)
56. Parker, L.E. Distributed Intelligence: Overview of the Field and its Application in Multi-Robot Systems. Journal of Physical Agents **2**(1) 5–14 (2008)
57. Passino, K.M. Biomimicry of Bacterial Foraging for Distributed Optimization and Control. IEEE Control Systems Magazine **22**(3) 52–67 (2002)
58. Pilcher, J.J., Nadler, E., and Busch, C. Effects of hot and cold temperature exposure on performance: A meta-analytic review. Ergonomics. **45(**10) 682-698 (2002)
59. Reimer, B., Mehler, B., Coughlin, J.F., Godfrey, K.M., and Tan, C. An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '09). ACM, New York, 115-118 (2009)
60. Simulation of Human Performance, **5** 469–498. Emerald Group Publishing (2004)
61. Roscoe, A.H. Assessing pilot workload. Why measure heart rate, HRV and respiration? Biological Psychology. **34** 259-287 (1992)
62. Russell, S.J., and Norvig, P. Artificial Intelligence: A Modern Approach, 2nd ed. Prentice Hall (2002)
63. Salvucci, D.D. Predicting the effects of in-car interface use on driver performance. Human-Computer Studies, **55** 85-107 (2001)
64. Scheutz, M., Schermerhorn, P., Kramer, J., and Anderson, D. First steps toward natural human-like HRI. Autonomous Robots. **22**(4) 411-423 (2007)

65. Scholtz, J. Theory and Evaluation of Human Robot Interactions. Proceedings of the 36th Hawaii International Conference on Systems Science, (2003)

66. See, J. E., Warm, J.S., Dember, W.N., and Howe, S.R. Vigilance and signal detection theory: An empirical evaluation of five measures in response bias. Human Factors. **39(**1) 14-29 (1997)

67. Silverman, B.G. Toward Realism in Human Performance Simulation. In: J. W. Ness, V. Tepe, and D. R. Ritzer (eds.) The Science and Simulation of Human Performance (Advances in Human Performance and Cognitive Engineering Research, Volume 5), Emerald Group Publishing Limited, 469-498 (2004)

68. Silverman, B.G., Johns, M., Cornwell, J., and O'Brien, K. Human Behavior Models for Agents in Simulators and Games: Part I: Enabling Science with PMFserv. Presence: Teleoperators and Virtual Environments **15**(2) 139–162 (2006)

69. Silverman, B.G., Pietrocola, D., Weyer, N., Weaver, R., Esomar, N., Might, R., and Chandrasekaran, D. NonKin Village: An Embeddable Training Game Generator for Learning Cultural Terrain and Sustainable Counter-Insurgent Operations. In: F. Dignum, J. Bradshaw, B. Silverman, W. van Doesburg (eds.) Agents for Games and Simulations, LNCS 5920, 135–154. Springer Berlin (2009)

70. Simmons, R., Singh, S., Hershberger, D., Ramos, J., and Smith, T. First Results in the Coordination of Heterogeneous Robots for Large-Scale Assembly. In: Experimental Robotics VII, LNCS 271, 323–332. Springer (2001)

71. "Simple Triage and Rapid Treatment (START)." Community Emergency Response Team Los Angeles. <http://www.cert-la.com/triage/start.htm>. Web. 15 Sept. 2010

72. Sjogren, P., and Banning, A. Pain, sedation and reaction time during long-term treatment of cancer patients with oral and epidural opiods. Pain. **39** 5-11 (1989)

73. Sternberg, S. Two operations in character recognition: Some evidence from reaction-time measurements. Perception and Psychophysics. **2** 45-53 (1967)

74. Stilwell, D.J., and Bay, J.S. Toward the Development of a Material Transport System using Swarms of Ant-like Robots. In: Proceedings of IEEE International Conference on Robotics and Automation, **1** 766–771. Atlanta, GA, USA (1993)

75. Stone, P., and Veloso, M. Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. Artificial Intelligence **110**(2) 241–273 (1999)

76. Sun, S.J., Lee, D.W., and Sim, K.B. Artificial Immune-Based Swarm Behaviors of Distributed Autonomous Robotic Systems. Proceedings of IEEE International Conference on Robotics and Automation, **4** 3993–3998 (2001)

77. Tang, F., and Parker, L.E. ASyMTRe: Automated Synthesis of Multi-Robot Task Solutions through Software Reconfiguration. Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA), 1501–1508. IEEE (2005)

78. Tang, F., and Parker, L.E. Distributed multi-robot coalitions through ASyMTRe-D. International Conference on Intelligent Robots and Systems (IROS), 2606–2613. IEEE (2005)

79. Tsalatsanis, A., Yalcin, A., and Valavanis, K.P. Optimized Task Allocation in Cooperative Robot Teams. 17th Mediterranean Conference on Control and Automation. 270–275 (2009)

80. van Lent, M., Silverman, B.G., McAlinden, R., O'brien, K., Probst, P., and Cornwell, J. Enhancing the Behavioral Fidelity of Synthetic Entities with Human Behavior Models, in Proc. of 13th Conference on Behavior Representation in Modeling and Simulation, (2004)

81. Vincente, K.J., Thornton, D.C., and Moray, N. Spectral analysis of sinus arrhythmia: a measure of mental effort. Human Factors. **29**(2) 171-182 (1987)

82. Vig, L., and Adams, J.A. Multi-Robot Coalition Formation. In: IEEE Transactions on Robotics, **22**(4) 637–649 (2006)

83. Vig, L., and Adams, J.A. Coalition Formation: From Software Agents to Robots. Journal of Intelligent and Robotic Systems **50**(1) 85–118 (2007)

84. Vincent, R., Fox, D., Ko, J., Konolige, K., Limketkai, B., Morisset, B., Ortiz, C., Schulz, D., and Stewart, B. Distributed multirobot exploration, mapping, and task allocation. Annals of Mathematics and Artificial Intelligence **52**(2–4) 229–255 (2008)

85. Vogler, David. "Raw Video Footage / WTC 9.11.01." 2001. http://davidvogler.com/91 Web. 31 Aug. 2010.

86. Wickens, C.D., Dixon, S., and Chang, D. Using interface models to predict performance in a multiple-task UAV environment – 2 UAVs. Technical Report for the Aviation Human Factors Division Institute of Aviation prepared for Micro Analysis and Design. (2003)

87. Wickens, C.D., Lee, J.D., Liu, Y., and Gordon-Becker, S.E. An Introduction to Human Factors Engineering, 2nd ed. Prentice Hall (2003)

88. Zambonelli, F., Jennings, N.R., and Wooldridge, M. Organisational Rules as an Abstraction for the Analysis and Design of Multi-Agent Systems. International Journal of Software Engineering and Knowledge Engineering **11**(3) 303–328 (2001)

89. Zhang, Y., and Parker, L.E. A General Information Quality Based Approach for Satisfying Sensor Constraints in Multirobot Tasks. IEEE International Conference on Robotics and Automation, 1452–1459 (2010)

90. Zhong, C. Modeling Humans as Peers and Supervisors in Computing Systems through Runtime Models. Ph.D. thesis, Kansas State University (2012)

91. Zlot, R., and Stentz, A. Market-based Multirobot Coordination for Complex Tasks. The International Journal of Robotics Research **25**(1) 73–101 (2006)